Models and Algorithms for Transient Queueing Congestion in Airline Hub and Spoke Networks

by

Michael Downes Peterson

A.B. Princeton University
(1986)
M.A. University of Cambridge
(1988)

Submitted to the Sloan School of Management
in Partial Fulfillment of
the Requirements for the Degree of
DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

Massachusetts Institute of Technology

September 1992

Signature of Author_____
Sloan School of Management
June 22, 1992

Certified by_____
Amedeo R. Odoni
Professor of Aeronautics and Astronautics and of Civil Engineering
Thesis Supervisor

Certified by_____
Dimitris J. Bertsimas
Associate Professor of Operations Research
Thesis Supervisor

Accepted by_____
James B. Orlin
Chairman, Sloan School Doctoral Program Committee

Models and Algorithms for Transient Queueing Congestion in Airline Hub and Spoke Networks

by

Michael Downes Peterson

Submitted to the Sloan School of Management
on June 22, 1992
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN MANAGEMENT

Abstract

This dissertation studies the relationship between the hub-and-spoke design in air transportation and the phenomenon of landing congestion. The problem of modeling this congestion is difficult because of the necessity of capturing transient rather than steady state behavior. To address the problem, we combine a deterministic treatment of arrivals with a model of capacity as a Markov or semi-Markov process, and from this we develop a computational approach for predicting transient queueing delays. In the special case where the demand rate is constant, we also develop an alternative method using a diffusion approximation adapted for this system. We provide computational results comparing the two approaches. 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 impossible. We explore in depth two questions of policy: schedule interference between the two principal carriers, and the likely effects of demand smoothing policies on queueing delays. In the final part of the thesis, we extend the analysis by developing two algorithms for congestion in a hub-and-spoke network. These algorithms are decomposition approaches which treat queues individually while approximating alterations in the downstream arrivals. Tests of the algorithms against a simple simulation model on a small network indicate that these approximations work fairly well, though they tend to underestimate the spreading of traffic which occurs as a result of delays.

Thesis Jointly Supervised by:

Amedeo R. Odoni	Professor of Aeronautics and Astronautics and of Civil Engineering
Dimitris J. Bertsimas	Associate Professor of Operations Research

Acknowledgments

I suppose that the dissertation is one of life's supreme educational experiences — it certainly has been for me. Like most forms of education, this one came with teachers, those whose advice guided me throughout the work. For their contributions to this part of my education, I owe a great debt of thanks to my two co-advisors, Amedeo Odoni and Dimitris Bertsimas, both of whom have provided me with advice and encouragement beyond the confines of queueing and airports. For their constant feedback throughout this project I am enormously grateful.

My thanks also go to the third member of my committee, Arnold Barnett, whose humor has leant me a sense of perspective during the hectic final portions of this work. Although I can never be like him, he is nevertheless a role model for me in the classroom. Richard Larson, Co-Director of the Operations Research Center at M.I.T., is another of my influential teachers, though he has not been directly involved in this work. Dick's emphasis on deriving research from real problems has significantly influenced my view on what operations research ought to be.

Many details of the thesis could not have been produced without the help of others. With respect to the issue of data in particular, I am indebted to American Airlines Decision Technologies, which provided data for Chapter 4 as well as a day of helpful discussion in July 1991. I single out Jim Diamond and Rick Dietert for their ongoing assistance. Thanks also to Bill Swan and Craig Hopperstadt at Boeing Computing Services for their advice in early stages of the thesis, and to Otis Welch of the F.A.A. Southwest Region and Katherine L. Houk of the Dallas-Fort Worth Airport Authority, who helped to supply capacity and traffic data. For assistance in computing and advice in de-bugging, I thank John Maglio, Rob Shumsky, and Yusin Lee. Special thanks to Vikas Sharma, who has endured a cramped office with me during these final months and who has been very generous in allowing me the use of his computer for stages of the writing. Along this same line, thanks to Krishnan for help with printing of the drafts.

My fellow students here have given me their encouragement, advice, and good humor. In particular, I would like to mention Pieter Klaassen, Menke Ubbens, Jari Kinaret, Richard Holme, Jens-Petter Falck, Steve Gilbert, Jim Walton, Rob Smith, and David Gebala. Outside the confines of M.I.T., I thank those whose friendship has meant a lot to me: John Young, Mohan Subramaniam, Felicia Brady, Susan Thornberg, Anja Bergman, and Kristi Cunningham. And of course I thank my family, whose support I do not acknowledge often enough.

Special thanks are reserved for Sarah, who has cheered me through an anxious and sometimes difficult year. This dissertation is dedicated to her.

This research was supported by a National Science Foundation Graduate Fellowship.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Hub-and-Spoke Phenomenon

Since deregulation was initiated in 1978, the U.S. airline industry has been characterized by turbulent change. The recent failures of two of the nation's oldest carriers, Pan American and Eastern, are the most recent testimony to this fact. Unable to adapt quickly enough, these carriers succumbed after a decade of fighting low-cost new entrants and established rivals who proved more innovative in a decade of airline innovations.

Among the most noticeable innovations of the 1980's was the development of extensive hub-and-spoke networks by the major carriers. Although American, United, and Delta (the three largest carriers in 1992) all operated hubs prior to the 80's, these operations bore little resemblance to the well-developed networks which exist today. In the deregulated environment, carriers have taken advantage of the freedoms granted to them by moving away from "linear" (point-to-point) systems to multi-hub hub-and-spoke networks. For example, American, United, and Delta today operate 6, 5, and 5 hubs respectively. The *failure* of Pan Am to develop a good hub-and-spoke network is widely regarded as one of the principal reasons for its demise.

There are strong economic motivations for developing hub-and-spoke networks.

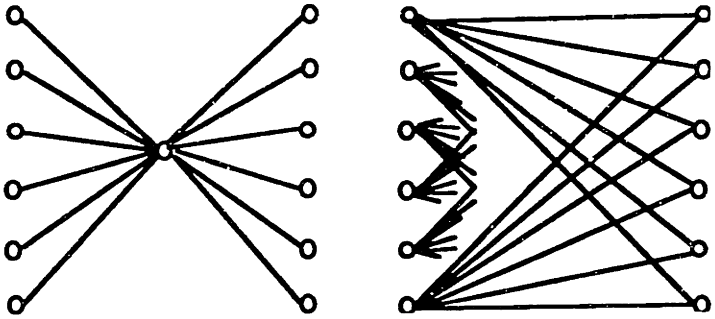


Figure 1.1: Illustration of connecting service advantage. In the hub system, the 36 east-west O-D (round-trip) pairs are served with just twelve flight legs. In the point-to-point system, all markets are served nonstop, but the number of flight legs necessary is three times as great.

A well-developed hub system consolidates demand in a large number of markets, allowing an airline to offer higher frequency of service to its customers and to take advantage of certain economies of scope and scale.[1] This demand consolidation allows an airline to serve profitably a city pair whose traffic level alone would not economically justify frequent service. It also reduces the uncertainty in the loads carried, allowing carriers to improve their average load factors.[2] In a more strategic vein, it is believed that the high frequencies of service to and from the hubs give carriers a degree of market power in setting their fares at these locations [11,36].

The inherent advantages of demand consolidation are easily illustrated in the hypothetical network of Figure 1.1. Here there are six "western" airports and six "eastern" airports. Ignoring trips within regions, the total number of round-trip markets is 36. In the case where a central hub is used to facilitate connecting passengers (left side of figure), a total of 12 flight legs are necessary to service demand while in the point-to-point network (right side of figure), 36 flight legs are

[1]Scale economies in aircraft operating costs imply that where a carrier can increase the size of aircraft with which it serves demand, it can reduce its expenses. Consolidation of markets allows the possibility of such activity. In connection with this, see [20]. Scope economies (i.e. those having to do with breadth rather than size of operations) are associated with the ground costs of servicing demand. See [26].

[2]The average load factor for an airline is defined as the ratio of revenue passenger miles (paying customers times number of miles flown) to available seat miles (seats offered times miles flown). It is a measure of how full the carrier's aircraft are on average.

necessary. The first network has an average traffic level per leg six times that of the second, creating higher frequencies, larger loads, or both, and resulting in cost and revenue benefits. But there are potential disadvantages in the design as well, as we indicate next.

1.2 Delay and the Hub-and-Spoke Network

While the economic advantages of the hub-and-spoke system for airlines are universally acknowledged, there is disagreement with respect to the system's benefits for consumers. A recent article in *The New York Times Magazine* 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.[3] Among the sources of this discontent are the belief that nonstop service has decreased with the advent of the hub-and-spoke design[4] and frustration with increased delays.

According to the legal director of the Aviation Consumer Action Committee,[5] delay is the principal reason for consumer dissatisfaction with deregulation. During the past fifteen years, congestion has become a fact of life at many major airports in this country and in Europe. 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 an entire 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.[7] Many consumers associate this increased delay with travel in a hub-and-spoke system. As a recent article in *The Economist* lamented,

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

[4] It is a common belief that the growth of hub systems has led to a reduction in nonstop service for passengers. However, there are no statistical studies supporting this conclusion. On the contrary, one recent study [6] finds that nonstop service has improved over the past decade, chiefly because there are now so many hubs with nonstop service to and from spoke cities.

[5] Cornish Hitchcock, interview in "Off Course", *N.Y.T.M.*, Sept. 1, 1991, page 14.

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

[7] *Winds of Change: Domestic Air Transport Since Deregulation*, Transportation Research Board National Research Council Special Report 230, Washington, D.C., September 1991, p.216.

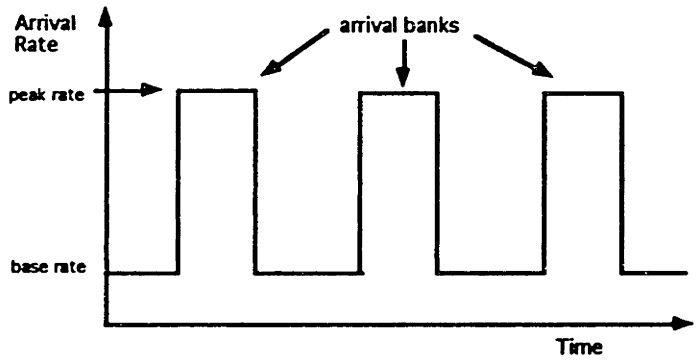


Figure 1.2: Flights concentrated into banks cause peaking of the landing demand rate throughout the day. During slack periods, the absence of bank arrivals reduces the demand rate.

> Frequent travellers are only too happy to describe, in paralysing detail, the horrors of flying in America. You must first negotiate thickets of fares and ticketing restrictions, designed to make sure that business travellers pay most. Then you will almost certainly have to shove your way through one of the big congested "hub" airports, where jets swarm in to swap passengers. Crowds and complications: such are the joys of flying these days.[8]

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 probably also played a role. Hubs are congested because they experience higher traffic levels (Figure 1.1). 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. Moreover, hub-and-spoke systems tend to concentrate major airport operations (landings and takeoffs) into short periods of time, placing further strain on capacity (Figure 1.2). This concentration of traffic can lead to queues for gates and runways, even when the overall daily load falls well within capacity limits. More-

[8] "Too Many Airlines", *The Economist*, October 19, 1991, page 13.

over, 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.

Operations research models like those presented in this dissertation have a clear role to play in improved airport planning. This role was well summarized by the National Transportation Research Board in its 1991 report "Winds of Change":

> The F.A.A. should emphasize research on simulation modeling of airport and airspace capacity and related research. Greater use of such techniques would lead to the establishment of performance measures that would help the F.A.A. make better use of existing airport and airspace capacity.[9]

The models developed here are intended as serious alternatives to simulation models, which are widely considered to be necessary because of the inherent complexity of airport queueing systems. As the following discussion indicates, this complexity arises chiefly from the need to model the *transient behavior* of airport queues.

1.3 Queues at Hub Airports

Queues develop at airports in numerous contexts (e.g. ticketing, baggage handling), but in this dissertation we are concerned with queues of *aircraft* which arise because of limited runway capacity. Figure 1.3 depicts the relevant set of operations. Incoming aircraft are seen as customers requiring service at a series of three stations: a landing runway, a gate, and a departure runway. Most of the work in this dissertation will focus on the landing queue, which is often the system bottleneck. However, our methods are largely applicable to the departure queue as well.

The airport queueing system has several characteristics which make it unfit for most traditional analyses, namely:

[9] *Winds of Change: Domestic Air Transport Since Deregulation*, p.299.

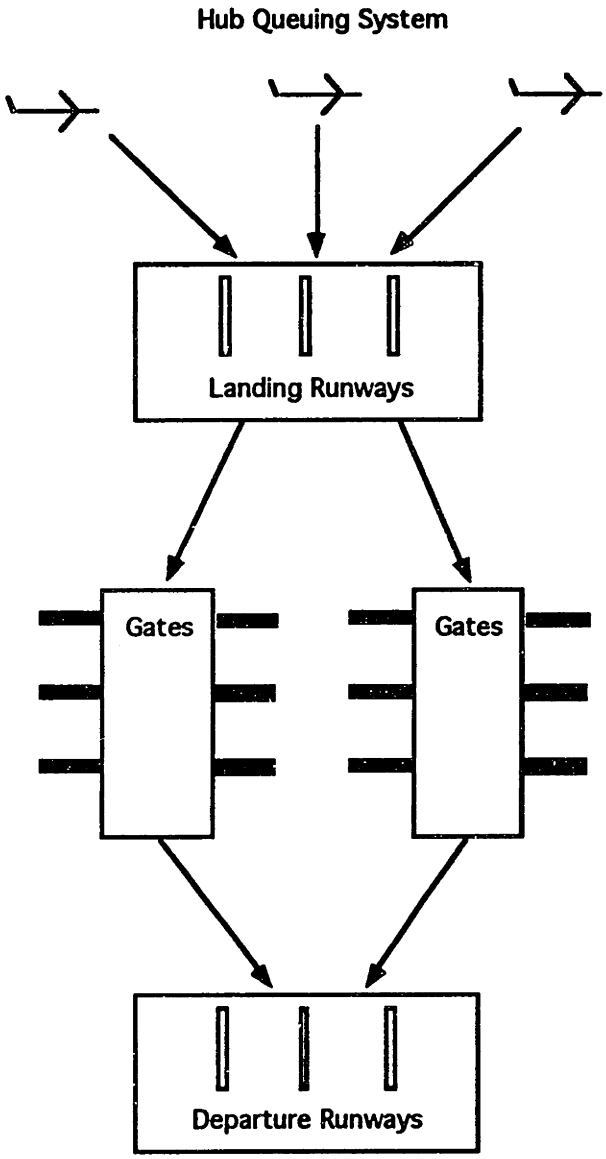


Figure 1.3: The aircraft queueing system at an airport

1. *Time variation in arrival rate.* As Figure 1.2 suggests, a hub airport is subject to a highly time-varying demand rate. Work comparing transient and steady state results for single queues [29,33] suggests that in such cases, the time necessary to reach a condition which is approximately "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. *Dependence of service times on weather.* For the landing and departure processes at an airport, service times are weather-dependent. Thus it is inappropriate to model service times as i.i.d.

3. *Customer dependence.* Aircraft in banks are not independent of one another. Because of connections between flights, an aircraft's time at the gate depends on the arrival times of other flights, as well as its own service time. Moreover, separation rules for large and small aircraft negate the assumption that consecutive landing service times are independent.

While these difficulties rule out most standard queueing approaches, they are not insurmountable. The simple deterministic models of Chapter 2, for example, can capture transient behavior, although they fail to model weather-dependence adequately. The models presented later in the thesis overcome this second difficulty by allowing capacity to vary as a stochastic process. Chapter 3 considers a single airport in isolation, but Chapter 5 goes further in accounting for dependencies between airports in a network. Modeling of dependencies between aircraft (item three) is more difficult and probably requires simulation. However, it is possible to gain considerable insight even while ignoring such dependencies.

1.4 Research to Date

In this section we review some of the literature relevant to the thesis. This work essentially falls into two categories: general studies of transient queue behavior and more specific studies of congestion at airports.

1.4.1 Studies of Transient Queueing Behavior

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 to the transient behavior of a single queue model the service and arrival processes as phase-type (i.e. Coxian) and attempt to solve the resulting forward Kolmogorov equations. Methods differ chiefly in the approach they take to solving these equations.

The most straightforward approach is to solve the Kolmogorov equations numerically. Gross and Harris [15] give a fairly thorough discussion of the competing methods — see especially their Section 7.3.2. Most of these methods 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 (or randomization). An early application of this method is given by Grassmann [14]. The essence of this method is that by uniformizing the underlying Markovian system, one substantially reduces the work necessary to obtain a solution. An explanation of this method is also found in Gross and Harris [15]. A third solution method due to Bertsimas and Nakazato [8] takes transforms of the Kolmogorov equations and then inverts these numerically to obtain the waiting time and queue length distributions. This is done for a system where the service and arrival distributions are mixed Erlang. A second paper [7] formulates the extension to general $GI/G/1$ systems as a Hilbert problem.

Odoni and Roth [29,33] 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 applications.

The above methods are "exact" in the sense that they seek solutions to the equations posed by a queueing model. An alternative approach is to approximate the queue length process by a continuous stochastic process, such as a Brownian motion. The latter obeys a certain partial differential equation, the solution of which, subject to initial conditions, yields the density for the (continuous) queue length. Details of this method, called the diffusion approximation, are given in our Chapter 3. The seminal papers on the subject are those of Iglehart and Whitt [18,19]. Kobayashi [24] applies the method to computer communication networks. Good summaries are found in Gelenbe and Mitrani [12] and Heyman and Sobel [17].

Our Chapter 5 deals with congestion in a network environment, a difficult problem. Successful modeling of general queueing network problems has been limited, and even under steady state assumptions, exact results have been obtained only for a relatively small class of problems which exhibit product-form solutions. The major reference in this field is that of Kelly [23]. Much research has focused on approximation methods, which may generally be divided into two types: diffusion approximations (just discussed) and decomposition methods. This second method consists of decomposing the network into its individual stations and approximating the departure process from each queue as a renewal process. The most prominent example of this kind of approach is the Queueing Network Analyzer [39], developed at Bell Labs for steady state analysis of non-product-form networks. We are not aware of any comparable approach for the transient problem.

1.4.2 Studies Related to Airport Congestion

Airport capacity and queueing studies have a history of over 30 years. The earliest work dates back to 1960 with the work of Blumstein [9] investigating the determinants of airport capacity. Airborne Instrument Laboratories, under contract to the F.A.A., developed the first handbook for estimating airport capacities [2,3], and thirteen years later the consulting company of Peat, Marwick, and Mitchell [30,31] published a new analysis based on simulation techniques. Newell [27] provides a thorough discussion of how airport geometry, flight rules, and weather conditions determine airport capacities. He claims, as we do, that standard queueing approaches are inadequate for airport queueing systems and argues instead for a deterministic approach similar to the one discussed at the end of our Chapter 2.

Two recent studies concern simulation approaches for estimating aircraft queueing delay. Abundo [1] considers the problem of queueing for landing at a single airport and provides some discussion of alternative methods. She proposes an $M(t)/E_k(t)/1$ model for the landing queue in combination with a simulation of the weather conditions in order to develop a capacity profile. She discusses the idea of modeling service capacity as a continuous-time Markov process but rejects this approach for two reasons. First, the state space necessary for an $M(t)/E_k(t)/1$ model together with a Markov weather process is computationally prohibitive. Second, she finds that for Boston's Logan airport, the subject of her case study, a Markov chain assumption for weather changes is not statistically warranted.

St. George [34] is concerned specifically with the issue of delay at hub airports. He uses a simple simulation model to estimate the delays expected to result from published airline schedules. His model 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. He provides a brief discussion of the effects of hubs on delay in comparing St. Louis (a hub) to Boston (a non-hub) but 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.

Recently, there has been a fair amount of work on the problem of flow control — how to alter in "real time" the flow of traffic between airports to reduce congestion. The central question is how long to hold aircraft on the ground prior to takeoff in order to avoid costly airborne holds at the congested destination airport. An introduction to the problem is given by Odoni [28]. Further work on the problem has been undertaken by Andreatta and Romanin-Jacur [4], Terrab [35], Richetta [32], and Vranas, Bertsimas, and Odoni [37,38].

1.5 Goal and Contribution of the Thesis

The goal of this thesis is the development of tractable and realistic models of congestion at a hub airport and within a hub-and-spoke system. By *realistic* we mean models which incorporate dependence between airports and which describe transient (not steady-state) behavior. As the discussion thus far has indicated, hub-and-spoke networks are central to the air transportation industry in this country and have definite implications for congestion. From the point of view of airport and airline planners, the modeling of these kinds of congestion delays is an issue of practical interest. Thus we believe that the first contribution of the thesis is the development of analytical models for single and multiple hub systems. As the case study of Chapter 4, in particular, is intended to show, the models we develop are of direct interest in addressing planning and policy questions. They also lead to a considerable amount of insight into the relationship between the hub-and-spoke system and the phenomenon of queueing delay.

We believe the thesis also makes contributions within the field of queueing theory. Queueing processes which require transient analysis are both numerous and understudied. Our single hub model is an entirely new method for describing the transient behavior of a certain type of single queue: one with simple probabilistic input and a service rate which changes according to a Markov or semi-Markov pro-

cess. Our explicit modeling of the dependence of the service rate on an external, stochastic phenomenon is, to the best of our knowledge, new within the literature on transient results in queueing theory. Our implementation of the diffusion approximation in Chapter 3 is the first instance of which we are aware of a real-world application with non-i.i.d. service times.

In Chapter 5 we introduce new approximation methods for estimating transient effects of queueing in air networks. While only partially developed here, these approaches suggest new directions for research in this rather difficult area of queueing theory. Potential applications are numerous in the fields of transportation, manufacturing, and communication.

1.6 Structure of the Thesis

The body of the dissertation is divided into four chapters. The first of these, Chapter 2, is introductory. Using a schematic model and deterministic analysis, we discuss several main features of hub congestion, in particular the effect of large demand peaks. We derive simple conditions for the queue to remain stable (i.e. not grow indefinitely) over time and introduce the notion of interference between carriers sharing a hub airport. We show that such interference may occur in situations where different carriers' banks are scheduled close together.

While these simple models serve as a useful introduction, their realism is severely compromised by the fact that they require deterministic specification of capacity. It is more desirable to have a model which explicitly takes into account the dependence of capacity on weather. This fact motivates our development of the computational approach introduced in Chapter 3, where we address the problem of predicting landing queue congestion as weather and capacity vary. With minor modifications, the method is applicable also to the departure queue.

Our model begins with the division of time into short increments, within which we assume demand and capacity to be constant. From interval to interval, we allow

capacity to vary according to either a Markov chain or a semi-Markov process, with a discrete number of states determined by weather conditions and runway configuration. We show how to calculate queue-length and waiting time moments for each time increment by using a simple recursive procedure and indicate how this procedure may be extended to the case where demand has a simple probabilistic (rather than deterministic) structure. Using the steady state probabilities for the capacity process, we obtain expected queue length and waiting time moments averaged over all initial conditions and sample paths.

In Chapter 4 we apply the recursive method in a case study of congestion at Dallas-Fort Worth Airport, one of the nation's busiest hubs. From weather data, we estimate parameters for Markov and semi-Markov models of capacity at DFW and give some discussion of how these compare. We then carry out a number of computational experiments which are addressed to relevant policy and planning questions. For example, 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 the average amount of delay. While rigorous validation of the model's results is difficult because of inadequacies in available delay data, informal analyses suggest that the model's estimates are reasonable. Our discussion illustrates the method's usefulness as a decision support tool.

In Chapter 5 we consider the more general problem of queueing delay in the environment of a hub-and-spoke *network*. In taking account of network effects, our task becomes more challenging: interactions between airports invalidate our earlier assumption that demand is known in advance. To address these difficulties, we propose two decomposition methods which estimate delays at individual airports according to the recursive procedures already developed and use these estimates to update downstream arrival rates. The first method simply updates arrivals according to the mean waiting times experienced by the aircraft, taking into account the slack present in the schedules. The second uses information about waiting time

variance to specify variability in the downstream effects. While both algorithms are applicable to general networks, we show how they are particularly well-suited for hub-and-spoke networks, where via a further simplification one can substantially reduce the size of the network which must be considered. In the second half of Chapter 5, we construct a hypothetical 2-hub network and test the two algorithms against a simple simulation procedure. Our results indicate that the approximation schemes work fairly well, though further improvement seems possible. The methods provide an effective way to study interaction between hubs in a network.

Chapter 6 summarizes the ideas of the thesis and suggests directions for further research.

Chapter 2

Congestion at One Hub: Introductory Models

While later chapters develop more precise computational models of congestion, the goal of this chapter is to develop an initial modeling approach. For the landing operations at a single hub airport, we consider a simple queueing model with a deterministic, time-varying demand rate and a constant service capacity. We focus on two issues, peaking of demand and schedule interference between carriers, addressing these in Sections 2.1 and 2.2 respectively. In our concluding remarks in Section 2.3, we place these results in perspective, commenting on the models' shortcomings and motivating the work of later chapters.

2.1 Model of a Single-Carrier Hub

In this section we consider a *monopoly* hub, an airport used by a number of carriers, only one of which (the "home carrier") operates its hub there. We classify each aircraft using this airport according to whether or not it belongs to the home carrier. Note that aircraft in the former category are organized into banks of arrivals and departures, while those in the latter arrive and depart in a less organized fashion throughout the day.

For simplicity of illustration, we assume that the home carrier's arrivals and

departures are organized into banks of lengths l and d respectively. Together, each landing bank and its subsequent departure bank define a *complex*. Within this complex, the two banks are separated in time by the *intra-complex slack time* s_1, while the time from the end of a departure bank until the start of the next arrival bank is the *inter-complex slack time* s_2 (see Figure 2.1). Within landing banks, the home carrier's aircraft have a constant landing demand rate λ_1; all other aircraft have a constant landing demand rate λ_2 throughout the day. Thus during peak periods the demand rate is the sum $\lambda_1 + \lambda_2$, while in slack periods it is merely λ_2. Corresponding to these demands we assume a constant landing capacity μ^l.

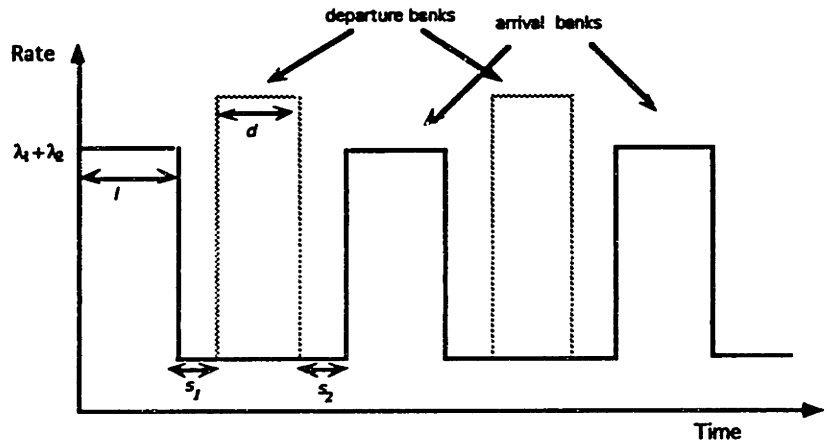
We are concerned with the transient behavior of the system and in particular with the effect of the peaking of demand on airport's ability to meet its schedule. Taking the simplest possible initial approach, we assume deterministic arrivals and a constant service rate. Under such assumptions, aircraft queues only develop during periods where demand exceeds capacity; thus to study the queue's behavior we shall assume

$$\lambda_2 < \mu^l < \lambda_1 + \lambda_2. \tag{2.1}$$

There are two questions we wish to address. The first concerns disruption of the connections schedule: under what conditions is the slack s_1 adequate to allow the completion of all landings in a given bank in time to begin the subsequent departure bank on schedule? Second and more important is the question of *stability:* are the slacks s_1 and s_2 adequate to allow the completion of one landing bank before the commencement of the next? When this latter condition fails, the queue for landing aircraft never reaches zero — the queueing situation becomes unstable.

These two questions may be addressed via the simple model. Consider first the issue of departure bank delays. Assume that each departure bank can only begin after the preceding arrival bank has landed and passengers and bags have connected.[1] Then our simple model implies the following condition for the slack

[1]This assumption is not realistic; in fact, the problem of deciding how long to hold departing aircraft at the gate is one which airlines face all the time. We make this assumption here in order to



$l \triangleq$ duration of a landing bank

$d \triangleq$ duration of a departure bank

$s_1 \triangleq$ intra-complex slack time

$s_2 \triangleq$ inter-complex slack time

$t_c \triangleq$ minimal time for passengers and luggage to connect between planes

$\lambda_1 \triangleq$ landing demand rate of home carrier

$\lambda_2 \triangleq$ landing demand rate of other carriers

$\mu^l \triangleq$ airport service rate for landings

Figure 2.1: One-carrier hub model

parameter s_1:

Proposition 2.1 *The arrival schedule is adequate to allow departing banks to commence without delay following their corresponding landing banks if and only if*

$$s_1 \geq t_c + \frac{(\lambda_1 + \lambda_2 - \mu^l) \times l}{\mu^l}. \tag{2.2}$$

PROOF:

At some point following time l (the end of the scheduled arrival bank), all arrivals in the bank will have landed. The difference between this time and the scheduled time is a period of duration

$$\frac{(\lambda_1 + \lambda_2 - \mu^l) \times l}{\mu^l}. \tag{2.3}$$

Thus departures can begin on schedule with all connections achieved if and only if (2.2) holds. □

In practice, typical values for t_c are about 20 minutes, and typical values for s_1 at busy hubs such as Dallas-Fort Worth are close to 25 minutes. Thus (2.2) suggests that even in good weather, minor passenger connection problems may arise frequently, though the delays are not likely to be severe. For example, for a large hub operating at a capacity of 90 landings per hour, a peak rate exceeding capacity by 10% translates into a queue of 9 aircraft and a delay of about 6 minutes in the start of the departure bank. Delays of this order of magnitude are commonly observed in practice.

While connection delay is clearly a problem, it does not constitute instability in the sense described above. The notion of instability is analogous to the queueing inequality $\rho > 1$, where ρ is the traffic intensity, the ratio of the average arrival rate to the average service rate. In our model, the natural expression for the this traffic intensity is the time average demand rate divided by the capacity:

$$\rho \triangleq \frac{1}{\mu^l}\left[\frac{l}{l + s_1 + d + s_2}(\lambda_1 + \lambda_2) + \frac{s_1 + d + s_2}{l + s_1 + d + s_2}\lambda_2\right]. \tag{2.4}$$

simplify our assessment of the effects of the schedule on aircraft delays, while acknowledging that there are further service complications involved with deciding these other questions.

In these terms, the condition $\rho > 1$ in fact constitutes queue instability (in the sense described above), as the next proposition shows.

Proposition 2.2 *Under the assumptions of the model, the landing queue remains stable if and only if*

$$\frac{l}{l + s_1 + d + s_2}\left(\frac{\lambda_1 + \lambda_2}{\mu^l}\right) + \frac{s_1 + d + s_2}{l + s_1 + d + s_2}\left(\frac{\lambda_2}{\mu^l}\right) \leq 1 \tag{2.5}$$

PROOF:

Consider the beginning of a peak landing period. At the end of the time scheduled for landings, the queue has grown by the amount

$$(\lambda_1 + \lambda_2 - \mu^l) \times l. \tag{2.6}$$

In the slack period which follows, the effective rate at which the queue is reduced is $\mu^l - \lambda_2$. It then follows that the ratio of (2.6) to this rate is the length of time beyond the peak period required for the queue to decrease by the amount which it grew during that period. If this length of time exceeds the time until the start of the next peak period, instability results; if not, the queue is stable. Simple algebra yields (2.5). □

Instability is a comparatively rare phenomenon, even at busy hubs. For such a hub, we typically have $s_1 + s_2 + d \approx 1$ hour and $l \approx 20$ minutes, in which case stability would require

$$\mu^l \geq (1/4)\lambda_1 + \lambda_2,$$

a fairly mild condition which we would expect to hold in all but the worst capacity circumstances.

Propositions 2.1 and 2.2 underscore the importance of the slack parameters s_1 and s_2, which allow the hub carrier to recover from the effects of delay. Increasing the slack increases the robustness of the schedule, though at the price of lower aircraft utilization in periods of adequate capacity.

2.2 A 2-Carrier Hub and the Notion of Interference

While the typical situation is for only one airline to operate a given location as a hub, there are exceptions to this rule. The most prominent example of a two-airline hub is Chicago's O'Hare Airport, where United operates its largest hub and American its second-largest after Dallas-Fort Worth. Until the demise of Eastern, that airline shared Atlanta with Delta. Delta shares a hub with American at Dallas.

The model of the previous section may be adapted to a 2-carrier situation. Consider Figure 2.2, which illustrates a general pattern of arrivals at a 2-carrier hub over time. These are now categorized into three types, for convenience numbered 1,2, and 3. Type 1 arrivals belong to the principal carrier at the hub (i.e. the carrier with the highest number of flights), type 2 arrivals belong to the other home carrier, and type 3 arrivals constitute all others.

We again make a number of simplifying assumptions. Type 3 arrivals are scheduled to arrive and depart throughout the day at uniform rate λ_3. Within time periods, arrivals of different types are scheduled uniformly. Peak demand rates are $\lambda_1 + \lambda_3$ during carrier 1's landing banks and $\lambda_2 + \lambda_3$ during carrier 2's. We also assume that the demands and service capacities satisfy

$$\lambda_1 + \lambda_3 > \mu^l$$

$$\lambda_2 + \lambda_3 > \mu^l$$

$$\lambda_3 < \mu^l$$

$$\lambda_1 \geq \lambda_2$$

and that takeoffs and landings utilize separate runways.

A notion of particular interest is that of interference. We define this to mean the situation where *the schedule of one carrier produces delays which are experienced by the other.* Figure 2.3 illustrates the idea. In the first part of the figure, the demand peaks of the two carriers are spread apart, giving a relatively greater chance that the delays produced by the congestion of one carrier do not affect the other. In other

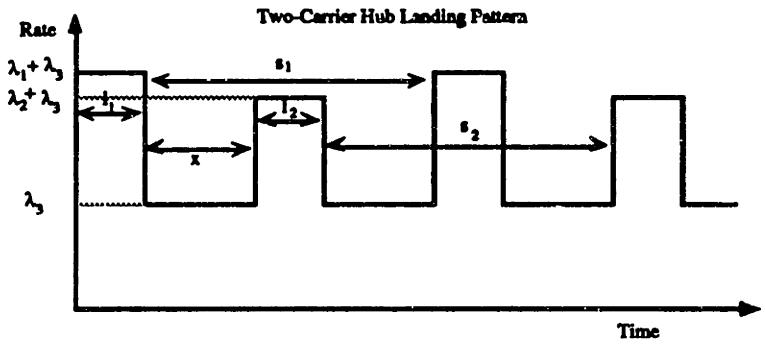


Figure 2.2: Two-carrier hub model

$l_1 \triangleq$ scheduled duration of each carrier 1 landing bank

$l_2 \triangleq$ scheduled duration of each carrier 2 landing bank

$\lambda_1 \triangleq$ carrier 1 landing demand rate

$\lambda_2 \triangleq$ carrier 2 landing demand rate

$\lambda_3 \triangleq$ landing demand rate for all other carriers

$\mu^l \triangleq$ landing service capacity (rate)

$s_1 \triangleq$ slack between banks for carrier 1

$s_2 \triangleq$ slack between banks for carrier 2

$x \triangleq$ scheduled time from end of carrier 1's bank until start of carrier 2's bank

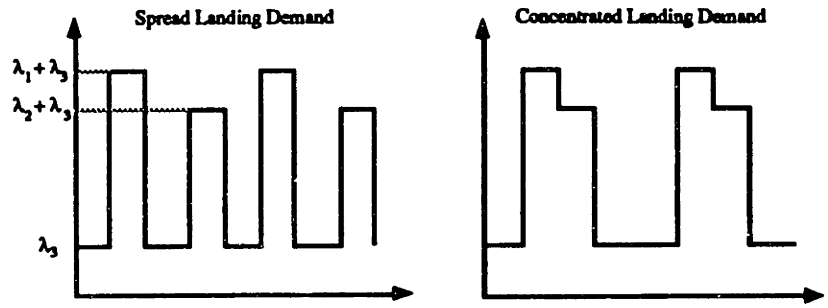


Figure 2.3: Alternate scenarios of arrival schedule at 2-carrier hub. The left side shows a schedule in which the two carriers have staggered banks, a schedule less prone to disruption. The right side shows a case where carrier 1's schedule is likely to disrupt that of carrier 2.

words, both carriers' banks have slack. The second half of the figure, in contrast, shows a situation where carrier 1 has its arrival banks positioned just prior to those of carrier 2. This kind of schedule is maximally disruptive to the latter, which must await the clearing of any queue created by carrier 1 before it can begin to land its aircraft.

The value of the parameter x is critical because it allocates the available slack between the two carriers and thus determines which is more likely to experience disruption of its schedule. A value of 0 for x, for example, gives a scheduled demand rate like the second half of Figure 2.2, favoring carrier 1. Conversely, a maximal x-value of $s_1 - l_2$ is most favorable to carrier 2.

We can use our simple deterministic model to analyze this situation further. Define the cycle length C by

$$C \triangleq l_1 + s_1 = l_2 + s_2.$$

The following result establishes a necessary and sufficient condition for stability of the landing queue.

Proposition 2.3 *In the above model of a 2-carrier hub, the landing queue remains stable over time if and only if*

$$(\lambda_1 + \lambda_3)(l_1/C) + (\lambda_2 + \lambda_3)(l_2/C) + \lambda_3(C - l_1 - l_2)/C \leq \mu^l. \qquad (2.7)$$

PROOF:

The proof is entirely analogous to that for Proposition 2.2. A necessary and sufficient condition for stability over a cycle is that the total number of landing aircraft is less than or equal to the total capacity:

$$l_1(\lambda_1 + \lambda_3) + l_2(\lambda_2 + \lambda_3) + (C - l_1 - l_2)(\lambda_3) \leq C\mu^l.$$

Dividing through by C yields the result. □

In theory, it is possible for one carrier or the other to experience persistent schedule disruption despite the fact that (2.7) holds. This phenomenon is a direct consequence of the value of x, the slack allocation parameter. Let us say that carrier 1 *delays* carrier 2 if the schedule is such that the initial arrivals of each carrier 2 bank are delayed. Similarly, carrier 2 delays carrier 1 if the initial arrivals of each of the latter's banks is delayed. We then have the following theorem.

Theorem 2.1 *Necessary and sufficient conditions for carrier 1 to delay carrier 2 while carrier 2 does not delay carrier 1 are that the queue is stable and that*

$$(\lambda_1 + \lambda_3)\,[l_1/(l_1 + x)] + \lambda_3\,[x/(l_1 + x)] > \mu^l. \tag{2.8}$$

Conversely, necessary and sufficient conditions for carrier 2 to delay carrier 1 while carrier 1 does not delay carrier 2 are that the queue is stable and that

$$(\lambda_2 + \lambda_3)\,[l_2/(s_1 - x)] + \lambda_3\,[s_1 - l_2 - x/(s_1 - x)] > \mu^l. \tag{2.9}$$

PROOF:

The proof for (2.8) is identical to that for (2.9), so it is sufficient to prove only the former. Suppose that the queue is stable and that (2.8) holds. Then over the period from the start of carrier 1's bank until the start of carrier 2's bank, the queue grows since average demand exceeds average capacity. Thus carrier 2 arrivals are delayed. But because the queueing process is stable, this queue must disappear by the start of carrier 1's next bank, for otherwise the queue would grow during each

cycle. Hence carrier 1 delays 2 but carrier 2 does not delay 1. Conversely, if carrier 1 is not delayed by carrier 2, then there is no queue at the start of carrier 1's bank. In order for carrier 2 to be delayed under these conditions, the condition (2.8) must hold. □

The result demonstrates the role of x in determining schedule disruption for the two carriers. For large values of x, it becomes increasingly likely that (2.9) holds, while smaller values favor (2.8). The important point is that *it is possible for one carrier to experience disruption while the other one operates on time.* The allocation of the available slack, while not usually done in a formal way, is decisive.

A "fair" schedule might have the separation between competing banks at the hub so that each carrier contributes equally to the chance of disruption for the other. Mathematically, this condition would be expressed as

$$\left[l_1(\lambda_1 + \lambda_3 - \mu^l)/(\mu^l - \lambda_3) - x\right]^+ = \left[l_2(\lambda_2 + \lambda_3 - \mu^l)/(\mu^l - \lambda_3) - (s_1 - l_2 - x)\right]^+ . \tag{2.10}$$

In practice, of course, no such formal planning takes place, and carriers are free to schedule as they wish (with the exception of slot-controlled airports). It is interesting to note that at one prominent 2-carrier hub, Dallas-Fort Worth, the schedule which results from this process may tend to favor one carrier (American) over another (Delta) in terms of queueing delays. We shall address this issue in the case study of Chapter 4.

2.3 Concluding Remarks

The schematic models of this chapter simplify the queueing problem to a greater degree than is necessary even for a deterministic approach. A more general time-varying demand rate $\lambda(t)$ can be accommodated within a deterministic scheme, and one could in theory specify a time-varying service rate $\mu(t)$ as well. The problem with the latter, however, is that unlike the demand rate, it is not easily specified in

advance or with certainty. One possibility is to employ such a model for a variety of capacity levels and examine the sensitivity of waiting times. However, such a model still ignores the possibility that capacity may change within the course of the operating day. To gain a thorough understanding, we must go further in specifying a stochastic capacity model. That challenge motivates the work of the next chapter.

Chapter 3

A Computational Approach to Congestion at a Hub

The present chapter is motivated by the need to develop a more powerful model of hub operations in order to improve our understanding of congestion at a hub airport. In particular, we consider the dependence of capacity on weather conditions and thereby address the major shortcoming of Chapter 2.

The model which we develop is intended as a strategic tool for airport and airline planning. Moreover, because the transient queueing environment is not unique to air transportation, our model is also motivated by potential applications in other fields such as manufacturing and communications. In its full generality, it represents a new approach for modeling queueing systems with time-varying arrival rates and state-dependent service times.

The chapter is organized as follows. In Section 3.1 we discuss the arrival and service operations for the landing queue at an airport and develop a model of capacity based on a semi-Markov process, which may be specialized to the simpler case of a Markov chain. In Section 3.2 we present a computational method for calculating queue length and waiting time moments for the landing process. The four parts of this discussion address, respectively, the basic recursions under deterministic demand (subsections 3.2.1 and 3.2.2), an extension to simple probabilistic demand (subsection 3.2.3), and the problem of taking averages over initial condi-

tions (subsection 3.2.4). In Section 3.3 we present computational results for a test case with constant demand. In Subsection 3.3.1 we show how the effects of initial conditions may be slow to diminish over time and how even with a constant demand rate, the time until equilibrium may be quite long. We also compare our results with those from simulation and examine in particular the empirical distribution of waiting times. In Subsections 3.3.2 and 3.3.3 we consider an alternative approach based upon a diffusion approximation of the queue-length process. While this kind of approximation has been applied to queueing systems where service times are independent and identically distributed, we believe our analysis to be the first such application to a queue where capacity varies according to a Markov chain. In Subsection 3.3.4 we compare the results of the diffusion approximation to those obtained from the recursive algorithm. While the results differ somewhat, the two methods seem to capture the same essential behavior.

3.1 Models of Demand and Capacity

In this section we describe our assumptions about the demand for aircraft landings and about airport service capacity.

As in Chapter 2, in this chapter we consider aircraft as customers utilizing a set of runways which together constitute a single server. We continue to treat the aircraft demand process as deterministic, assuming that aircraft follow *schedules* and do not just demand to land "at random." In practice, of course, this assumption is not strictly valid: arrival schedules contain elements of uncertainty because of earlier delays. We adopt the deterministic assumption in our initial approach and indicate in Section 3.2.3 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.

We summarize our view of demand in the following assumption:

Assumption 3.1 (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.

Consider now the service process. 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 chiefly 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 and thus constitutes a stochastic process with a discrete number of potential states.

In this thesis we employ two alternative models of such a process, one based on a Markov chain and one based on a semi-Markov process. In both cases, the states are the different capacities attainable by the airport under different weather scenarios and runway configurations. In the Markov model, capacity remains in a given state for a length of time Δt and then undergoes a transition, with self-transitions possible. Thus the number of periods for which the capacity remains in a given state (the holding time) is a random variable having geometric distribution. In the semi-Markov case, the holding time is allowed to have a general (discrete) distribution which may depend upon the state.[1]

In the computational approach which we develop shortly, we employ the more general semi-Markov formulation, from which one can easily deduce the results for

[1]This is not quite the most general form of a semi-Markov process. In its most general form, a semi-Markov process on a finite state space has a transition matrix $\mathcal{P} = \{p_{ij}\}$. When the process is in state i, then conditional on the next transition being to state j, the amount of time spent in i is a random variable with distribution $F_{ij}(x) = Pr\{T_i \leq x \mid i \rightarrow j\}$. In the model proposed here, the length of time in any state is assumed to be independent of the next state, so that $F_{ij}(x) = F_i(x) \forall j$. This does not seem to be a bad assumption with respect to weather changes.

the Markov case. However, the greater computational simplicity of the latter will prove to be of value in the case study of Chapter 4. We formalize our assumptions concerning the service process as follows:

Assumption 3.2 (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. These capacities are scaled according to the interval length Δt and obey the relationship*

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

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}.

3.2 An Algorithmic Approach

Assumptions 3.1 and 3.2 describe the arrival and service processes for this queueing system. We now develop a computational method for describing its transient behavior. To do this, 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 $i = 1, \ldots, S$. This fact suggests that queue statistics may be obtained by a recursive procedure. The next two subsections give such recursions for the queue-length and waiting time moments, respectively.

3.2.1 A Recursive Procedure for Transient Queue Lengths

To obtain queue-length moments, we begin by establishing a Markov chain within the semi-Markov process. We enlarge the state space to be $\{i, m\}$, where i is the current capacity state and m the age (in intervals) of that state, and we define the following random variables:

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

The following proposition describes the evolution of the capacity-age process.

Proposition 3.1 *The combined capacity-age process with state space (i, m) is Markov. For a given state (i, m), the transition probabilities are given by*

$$
\begin{aligned}
\tilde{p}_{ij}(m) \triangleq \Pr((i,m) \to (j,1)) &= \Pr[T_i = m \mid T_i \geq m] p_{ij} \quad j \neq i \\
\tilde{p}_{ii}(m) \triangleq \Pr((i,m) \to (i,m+1)) &= \Pr[T_i \geq m+1 \mid T_i \geq m] \qquad (3.1)
\end{aligned}
$$

PROOF:

For a given capacity and age (i, m), the possible states at the end of the next interval of time are $(j, 1)$ for all $j \neq i$ plus the state $(i, m+1)$. The probabilities of these transitions are clearly given by the formula (3.1), and since these transition probabilities, conditional on the state, are independent of prior history, the process is Markov. □

To define the recursive procedure for queue length, we introduce the notation

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

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

$$q_{\max}(k,i) = [q_{\max}(k-1) + \lambda_k - \mu_i]^+ \tag{3.2}$$

where

$$q_{\max}(k) \triangleq \max_i q_{\max}(k,i).$$

Let x^+ denote $\max(x,0)$. Then the recursion for expected queue lengths is given in the following theorem.

Theorem 3.1 *The functions $Q_k(l,i,m,q)$ obey the recursive relationship*

$$\begin{aligned} Q_k(l,i,m,q) \;=\; & \sum_{j\neq i} \tilde{p}_{ij}(m) Q_k\left(l{+}1, j, 1, (q+\lambda_{l+1}-\mu_j)^+\right) + \\ & \tilde{p}_{ii}(m) Q_k\left(l{+}1, i, m{+}1, (q+\lambda_{l+1}-\mu_i)^+\right) \end{aligned} \tag{3.3}$$

with boundary condition $Q_k(k,\cdot,\cdot,q) \equiv q$.

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 (3.3) now follows. □

The goal of the recursion (3.3) is to compute the values $Q_k(0,i,m,0)$ for all values of i, k, and m. These are expected future queue lengths conditional on the capacity state at the start of the day and on the forecasted demand. They

constitute a statistical prediction of how the queue is expected to behave, given initial conditions.

The recursion is not limited to finding first moments. Second moments (and hence variances) can also be found via a similar type of procedure. Define the second moment of the expected queue length by the shorthand

$$Q_k^2(l,i,m,q) \triangleq E\left[Q_k^2 \,|\, Q_l = q, C_l = i, A_l = m\right].$$

The previous conditional probability argument proves the following theorem for the second moments of queue length.

Theorem 3.2 *The functions $Q_k^2(l,i,m,q)$ obey the recursive relationship*

$$\begin{aligned} Q_k^2(l,i,m,q) \;=\; & \sum_{j\neq i} \tilde{p}_{ij}(m) Q_k^2\left(l+1, j, 1, (q+\lambda_{l+1}-\mu_j)^+\right) + \\ & \tilde{p}_{ii}(m) Q_k^2\left(l+1, i, m+1, (q+\lambda_{l+1}-\mu_i)^+\right) \end{aligned} \tag{3.4}$$

with boundary condition $Q_k^2(k,\cdot,\cdot,q) \equiv q^2$. □

Theorems 3.1 and 3.2 imply the algorithm for the queue length process given in Figure 3.1. The computational complexity and memory requirements of this algorithm are naturally of interest and are addressed in the following theorem.

Theorem 3.3 *The memory requirement for the semi-Markov queue length 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

$$2 \times S \times M \times \sum_{k=1}^{K} \sum_{l<k} q_{\max}(l). \tag{3.5}$$

Algorithm for Queue Length Moments

Boundary Condition:
For $k = 1$ to K
For $i = 1$ to S
For $m = 1$ to M
For $q = 0$ to $q_{\max}(k, c)$
$Q_k(k, i, m, q) = q$
$Q_k^2(k, i, m, q) = q^2$

Main Body:
For $k = 1$ to K
For $l = k-1$ down to 0
For $i = 1$ to S
For $m = 1$ to M
For $q = 0$ to $q_{\max}(l, c)$

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

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

END.

Figure 3.1: Recursive algorithm for queue length moments conditional on initial capacity and age conditions

Within iteration k, however, the algorithm needs only to store four values at a time, $Q_k(l,i,m,q)$ and $Q_k(l+1,i,m,q)$ for the first moment, $Q_k^2(l+1,i,m,q)$ and $Q_k^2(l+1,i,m,q)$ for the second. Thus since $q_{\max}(l) \leq Q_{\max}$ the memory requirement is $O(SKMQ_{\max})$. To calculate each table entry requires $O(S)$ time. Therefore the overall run time has complexity $O(S^2K^2MQ_{\max})$. □

The theorem indicates that the speed (and hence the practicality) of the computation rests on the relative sizes of K, M, and $Q_{\max}$, since S is very small (≈ 5). A full operating day is twenty hours at most ($K=80$), with typical values for $Q_{\max}$ in the range of 200. 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, it can be that above a certain value of m, the transition probabilities $\tilde{p}_{ij}(m)$ remain relatively constant over m. In rough terms, this means that while the holding time distributions may not be geometric, their tails might look approximately geometric. If 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 chapter, for example, a value of M as low as 20 proves adequate.

Obviously, substantial computational savings are available through reduction of M. 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. Run time and memory requirements are correspondingly reduced. These savings motivate consideration of both Markov and semi-Markov models in Chapter 4.

3.2.2 Computing Waiting Times

The previous discussion dealt with the evolution of the queue length process. This subsection addresses the question of waiting times. Let W_k be the waiting time for an aircraft scheduled to land at the end of the kth interval (i.e. at time $k\Delta t$). As for the queue length process, we define

$$\mathcal{W}_k(l,i,m,q) \triangleq E\left[W_k \,|\, Q_l = q, C_l = i, A_l = m\right].$$

We can establish a recursion for mean waiting times in a fashion similar to that for queue lengths. The main part of this recursion is contained in the next theorem. The proof, identical to that of Theorem 3.1, is omitted.

Theorem 3.4 *The functions $\mathcal{W}_k(l,i,m,q)$ obey the recursive relation (for $l < k$)*

$$\begin{aligned}\mathcal{W}_k(l,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\left[\mathcal{W}_k\left(l{+}1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \\ &\quad \tilde{p}_{ii}(m)\mathcal{W}_k\left(l{+}1,i,m{+}1,(q+\lambda_{l+1}-\mu_i)^+\right) \qquad (3.6)\end{aligned}$$

□

The complication with waiting times (as opposed to queue lengths) occurs at the boundary $l = k$. Let the notation $(a \wedge b)$ denote $\min(a,b)$. 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 3.5 *The functions $\mathcal{W}_k(k,i,m,q)$ obey the recursion*

$$\begin{aligned}\mathcal{W}_k(k,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\left[\left(\frac{q}{\mu_j}\wedge 1\right) + \mathcal{W}_k(k,j,1,(q-\mu_j)^+)\right] + \\ &\quad \tilde{p}_{ii}(m)\left[\left(\frac{q}{\mu_i}\wedge 1\right) + \mathcal{W}_k(k,i,m{+}1,(q-\mu_i)^+)\right] \qquad (3.7)\end{aligned}$$

where $\mathcal{W}_k(k,\cdot,\cdot,0) \equiv 0$.

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:

$$[W_k \mid \mathcal{I}] = [W'_k + W''_k \mid \mathcal{I}]. \tag{3.8}$$

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

$$\{Q_l = q, C_l = i, A_l = m\}.$$

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

$$[W'_k \mid \mathcal{I}, C_{k+1} = j] = \min(q/\mu_j, 1). \tag{3.9}$$

This follows because during the interval $k+1$ the queue in front of our 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 get 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

$$\begin{aligned}
[W''_k \mid \mathcal{I}, C_{k+1} = j] &\sim [W_k \mid Q_k = (q-\mu_j)^+, C_k = j, A_k = 1]\,,\ j \neq i \\
[W''_k \mid \mathcal{I}, C_{k+1} = i] &\sim [W_k \mid Q_k = (q-\mu_i)^+, C_k = i, A_k = m+1]\,.
\end{aligned} \tag{3.10}$$

Taking expectations of (3.9) and (3.10) and un-conditioning on $C_{k+1} = j$ yields the result. □

As with the queue lengths, the result of Theorems 3.4 and 3.5 are values for the expressions

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

These are expectations of waiting times at the end of each interval, based on given initial conditions.

To obtain second moments, we define the functions

$$\mathcal{W}_k^2(l,i,m,q) \triangleq E\left[W_k^2 \,|\, Q_l = q, C_l = i, A_l = m\right].$$

The part of the recursion for $l < k$ is the same as with the first moment. Proof is omitted.

Theorem 3.6 *The functions $\mathcal{W}_k^2(l,i,m,q)$ obey the recursive relation*

$$\begin{aligned}\mathcal{W}_k^2(l,i,m,q) \;=\; & \sum_{j\neq i} \tilde{p}_{ij}(m)\left[\mathcal{W}_k^2\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \\ & \tilde{p}_{ii}(m)\left[\mathcal{W}_k^2\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)\right] \qquad (3.11)\end{aligned}$$

□

Once again, there is a recursion within the major recursion to establish behavior at the boundary. This is described in the following theorem.

Theorem 3.7 *The functions $\mathcal{W}_k^2(k,i,m,q)$ obey the recursive boundary condition*

$$\begin{aligned}&\mathcal{W}_k^2(k,i,m,q) = \\ &\sum_{j\neq i} \tilde{p}_{ij}(m)\left[\left(\frac{q}{\mu_j}\wedge 1\right)^2 + 2\left(\frac{q}{\mu_j}\wedge 1\right)\mathcal{W}_k\left(k,j,1,(q-\mu_j)^+\right) + \mathcal{W}_k^2\left(k,j,1,(q-\mu_j)^+\right)\right] + \\ &\tilde{p}_{ii}(m)\left[\left(\frac{q}{\mu_i}\wedge 1\right)^2 + 2\left(\frac{q}{\mu_i}\wedge 1\right)\mathcal{W}_k\left(k,i,m+1,(q-\mu_i)^+\right) + \mathcal{W}_k^2\left(k,i,m+1,(q-\mu_i)^+\right)\right] \qquad (3.12)\end{aligned}$$

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

$$\{Q_l = q, C_l = i, A_l = m\}.$$

Using (3.8) we write $E[W_k^2 \mid Q_k = q, C_k = i, A_k = m]$

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

The final equality is a consequence of (3.9). The result (3.12) now follows from (3.10). □

Theorems 3.4 - 3.7 imply the algorithm for computing waiting time moments given in Figure 3.2. Memory requirements and running time complexity are the same as for the earlier queue length algorithm. We state this formally as our next theorem.

Theorem 3.8 *The memory requirement for the waiting time algorithm is* $O(SKMQ_{\max})$ *and the running time is* $O(S^2K^2MQ_{\max})$. □

It should be clear to the reader that the recursive approach of the queue length and waiting time algorithms could be used to obtain still higher moments, or indeed to recover 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

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

Algorithm for Waiting Time Moments

Boundary Condition I:

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

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

Boundary Condition II:

For $k = 1$ to K

For $i = 1$ to S

For $m = 1$ to M

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

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

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

Body:

For $k = 1$ to K

For $l = k-1$ down to 0

For $i = 1$ to S

For $m = 1$ to M

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

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

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

END.

Figure 3.2: Recursive algorithm for waiting time moments conditional on initial capacity and age conditions

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

3.2.3 Extension to a Simple Probabilistic Arrival Stream

The recursions for queue length and waiting time moments presented thus far are appropriate when the input stream is well approximated as a deterministic flow. This deterministic assumption is justified because aircraft are deliberately scheduled into their landing slots. On the other hand, congestion and other sources of 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, where delays at upstream airports affect others' arrival streams, it may be important to take account of this uncertainty. For this reason, we next introduce a straightforward extension of our approach which allows for a simple probabilistic structure in the demand process.

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. (3.3))

$$\begin{aligned} Q_k(l,i,m,q) &= \sum_{r=1}^{R} \gamma_{l+1}^r \left[\tilde{p}_{ii}(m) Q_k\left(l+1, i, m+1, (q+\lambda_{l+1}^r - \mu_i)^+\right) + \right. \\ &\qquad \left. \sum_{j \neq i} \tilde{p}_{ij}(m) Q_k\left(l+1, j, 1, (q+\lambda_{l+1}^r - \mu_j)^+\right) \right] \end{aligned} \tag{3.13}$$

with boundary condition $Q_k(k,\cdot,\cdot,q) \equiv q$. Similarly, for waiting times we have

$$\begin{aligned} W_k(l,i,m,q) &= \sum_{r=1}^{R} \gamma_{l+1}^r \left[\tilde{p}_{ii}(m) W_k\left(l+1, i, m+1, (q+\lambda_{l+1}^r - \mu_i)^+\right) + \right. \end{aligned}$$

$$\sum_{j \neq i} \tilde{p}_{ij}(m) \mathcal{W}_k \left(l+1, j, 1, (q + \lambda^r_{l+1} - \mu_j)^+\right) \Bigg] \quad (3.14)$$

Clearly, these additions to the algorithm multiply the running time by a factor R. Note that the method treats arrival rates in different periods as independent. That is, the random variables $\{\Lambda_k\}$ are 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 will be of value in Chapter 5, where the problem of congestion in the network is addressed.

3.2.4 Averaging Over Initial Conditions

The recursions implied by Theorems 3.1, 3.2, 3.4, and 3.6 obtain moments conditional on the starting state at the beginning of the day. 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

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

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

$$\overline{\mathcal{Q}}_k = \sum_{i,m} \pi(i, m) \mathcal{Q}_k(0, i, m, 0), \quad (3.15)$$

while the corresponding mean waiting time is

$$\overline{\mathcal{W}}_k = \sum_{i,m} \pi(i, m) \mathcal{W}_k(0, i, m, 0). \quad (3.16)$$

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

$$\pi = \pi \tilde{\mathbf{P}}, \quad (3.17)$$

where $\tilde{\mathbf{P}}$ is the full set of transition probabilities. A general linear system of this type would imply $S \times M$ equations and could be solved using Gaussian elimination in $O(S^3M^3)$ time. However, because of the special structure of the state space (see Figure 3.3), the solution to (3.17) can be obtained by solving a system of only S linear equations. This fact is shown in Theorem 3.9, which in turn depends on the following three propositions.

Proposition 3.2 *The steady state probabilities $\pi(i,m)$ for $m = 2,\ldots,M-1$ may be written in terms of the steady state probabilities $\pi(i,1)$ as*

$$\pi(i,m) = \pi(i,1) \prod_{k=1}^{m-1} \tilde{p}_{ii}(k) \tag{3.18}$$

PROOF:

Note from the diagram that for $m = 1,\ldots,M-1$, each state has only one entry point. Hence

$$\pi(i,m) = \pi(i,m-1)\tilde{p}_{ii}(m-1) \quad \text{for } m = 2,\ldots,M-1.$$

Successive substitution yields equation (3.18). □

Proposition 3.3 *The steady state probabilities $\pi(i,M)$ may be written in terms of the steady state probabilities $\pi(i,1)$ as*

$$\pi(i,M) = \pi(i,1)/(1-\tilde{p}_{ii}(M)) \prod_{k=1}^{M-1} \tilde{p}_{ii}(k) \tag{3.19}$$

PROOF:

For the states (i,M) we have the expression

$$\pi(i,M) = \pi(i,M-1)\tilde{p}_{ii}(M-1)/(1-\tilde{p}_{ii}(M)).$$

Repeated substitution yields equation (3.19). □

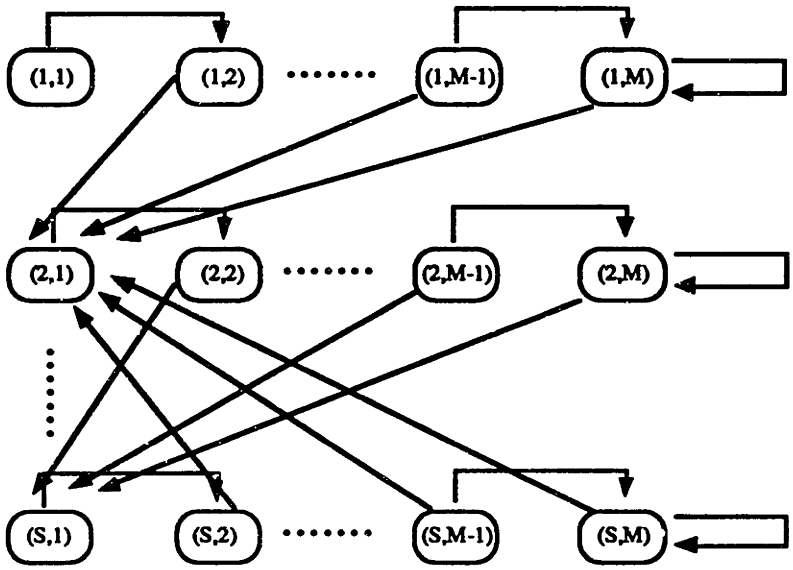


Figure 3.3: Diagram of the extended state space (i, m) for the semi-Markov model. Two kinds of transitions exist: those which keep the process in the same capacity (rightward arrows) and those which take the process to a different capacity (leftward arrows). The nodes labelled (i, M) correspond to states where the capacity s has prevailed for M *or more* periods. These states alone have the potential for self-transitions.

Proposition 3.4 *The steady state probabilities $\pi(i, m)$ must satisfy the system of linear equations*

$$\pi(i, 1) = \sum_{j \neq i} \sum_{m=1}^{M} \pi(j, m) \tilde{p}_{ji}(m). \tag{3.20}$$

PROOF:

The states $\pi(i, 1)$ can be entered from all other states in the system. The equations (3.20) are simply the standard balance equations for these states in the Markov chain (i_k, m_k). □

As a result of the simplifications implied in Propositions 3.2, 3.3, and 3.4, the problem of finding the steady state probabilities reduces to that of solving for the S unknowns $\pi(1, 1), \pi(2, 1), \ldots, \pi(S, 1)$.

Theorem 3.9 *The steady state probabilities $\pi(i,1)$ are the unique solution to the set of S linear equations*

$i = 1, \ldots, S-1$:

$$\sum_{j \neq i} \pi(j,1) \left[\sum_{m=1}^{M-1} \left(\tilde{p}_{ji}(m) \prod_{k=1}^{m} \tilde{p}_{jj}(k) \right) + \frac{\tilde{p}_{ji}(M)}{(1-\tilde{p}_{jj}(M))} \prod_{k=1}^{M-1} \tilde{p}_{jj}(k) \right] = \pi(i,1) \tag{3.21}$$

$$\sum_{i=1}^{S} \left[\pi(i,1) \left(\sum_{m=1}^{M-1} \prod_{k=1}^{m-1} \tilde{p}_{ii}(k) + \frac{1}{(1-\tilde{p}_{ii}(M))} \prod_{k=1}^{M-1} \tilde{p}_{ii}(k) \right) \right] = 1 \tag{3.22}$$

PROOF:

Propositions 3.2, 3.3, and 3.4 establish the necessity of equations (3.21). The Markov chain (i,m) is clearly irreducible and aperiodic, so the rank of this linear system is $S-1$. The normalizing condition (3.22) thus ensures a unique solution. □

The significance of this theorem is that 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 (3.21) and (3.22) for the probabilities $\pi(i,1)$ and then uses the relations (3.18) and (3.19) to solve for the others.

3.3 Queueing Theoretic Aspects

In this section we consider the proposed algorithm in a queueing theoretic context.

3.3.1 Comparison with Simulation

As a departure point for discussion, consider the results of a test run of the recursive algorithm for a constant arrival rate scenario: $\lambda = 68$, $\mu_{\min} = 50$, $\mu_{\max} = 112$, $\rho \approx 0.9$. In this test run, we employ the Markov chain version of the model.

Figure 3.4 depicts the mean queue length over time calculated by the Markov model. The three curves correspond to different starting conditions (lowest capacity,

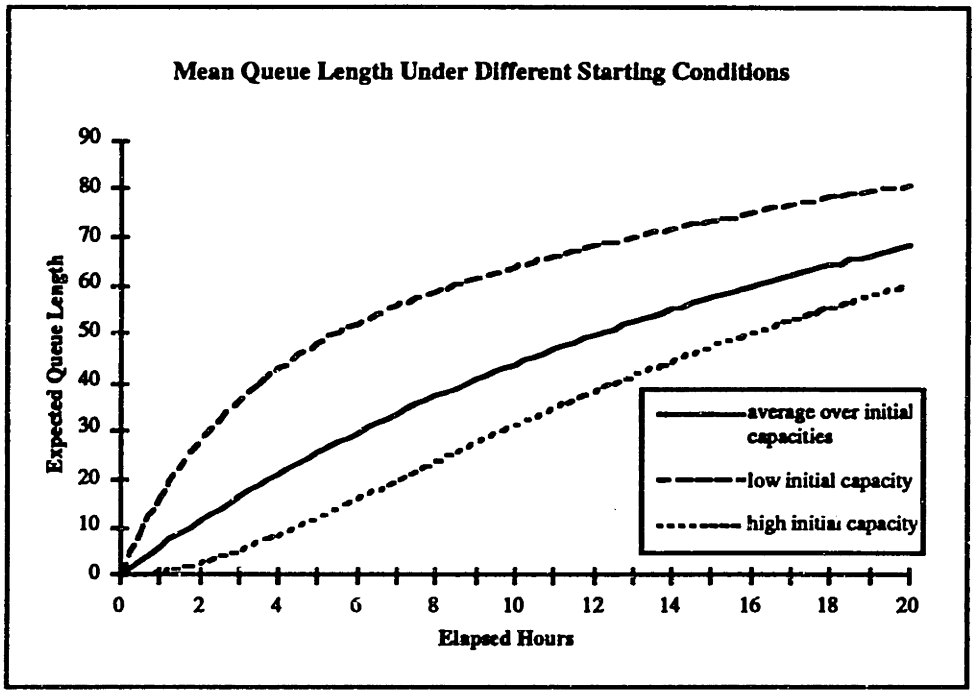


Figure 3.4: Comparison of expected queue length predicted by recursion under high and low capacity initial conditions and averaged over all initial conditions

highest capacity, and average over all capacities). From the figure, it is evident that the effects of the initial conditions do not wear off quickly over time, although the curves are obviously converging. Apparently, for these parameters the system is slow to reach steady state, a fact suggested by the positive slope of all three curves at $t = 20$ hours. The figure is evidence of the inappropriateness of steady state analysis, even for this constant demand system. Odoni and Roth [29] have indicated that the relaxation time for queues with i.i.d. service times can be quite large, especially for heavy traffic conditions. We conjecture that in our case, the relaxation time is still longer because of the grouping of services within intervals (i.e. all service times within an interval are perfectly correlated) as well as the correlation between successive intervals implied by the Markov chain.

As a check on our algorithm, we consider a simulation procedure in which we choose each period's capacity in Monte Carlo fashion according to the same Markov

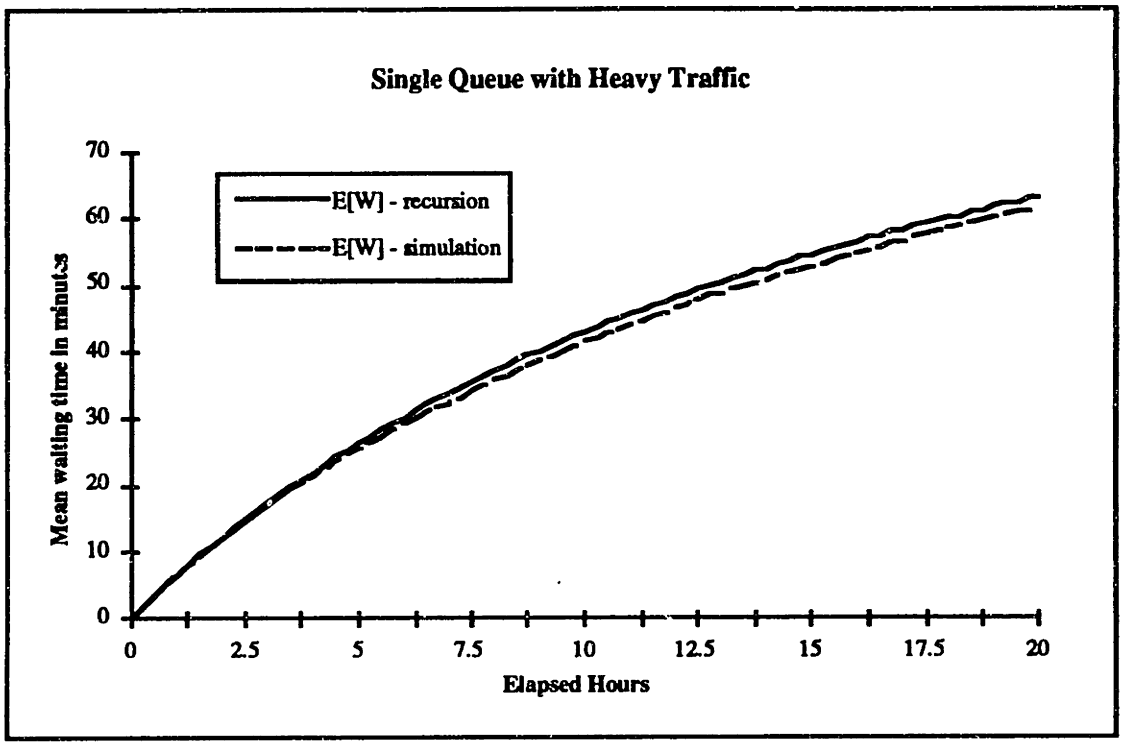


Figure 3.5: Comparison of mean waiting times estimated by simulation and by the recursive algorithm for a single queue with constant demand and heavy traffic

chain, trace the resulting changes in the queue, and then take averages of the resulting sample paths over different simulation runs. Our results are illustrated in Figures 3.5 and 3.6. Figure 3.5, based on average waiting times from 5000 simulation runs, indicates that the simulated mean values closely agree with those obtained from the recursion. The slight under-estimation which the simulation gives suggests that the tail occurrences for the waiting 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, we might expect them to have a noticeable effect on the standard deviations; Figure 3.6, which plots the latter for both the recursion and the simulation, confirms this.

Unfortunately, there is no easy way to rectify the sampling procedure to correct

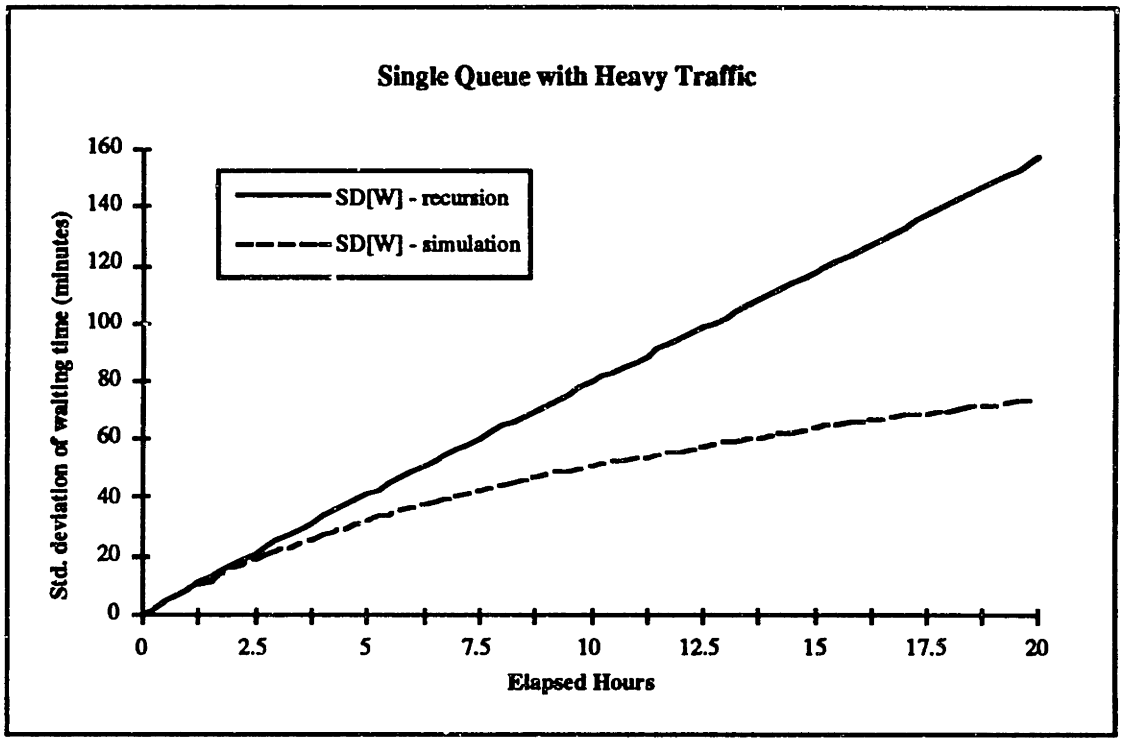


Figure 3.6: Comparison of standard deviations of waiting time estimated by simulation and by the recursive algorithm for a single queue with constant demand and heavy traffic

for this phenomenon. The waiting time at any given period k is a complicated function of the k capacities preceding it. Standard variance reduction techniques [16] such as stratified sampling and importance sampling are not readily applied to this multi-dimensional case.

As a final interesting use of simulation, we explore the question of waiting time distributions. Suppose 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 can obtain histograms for the waiting times for each period, like the one illustrated in Figure 3.7. Note the presence of a substantial probability mass at the minimum value (in this case, 0). Values above this minimum seem to

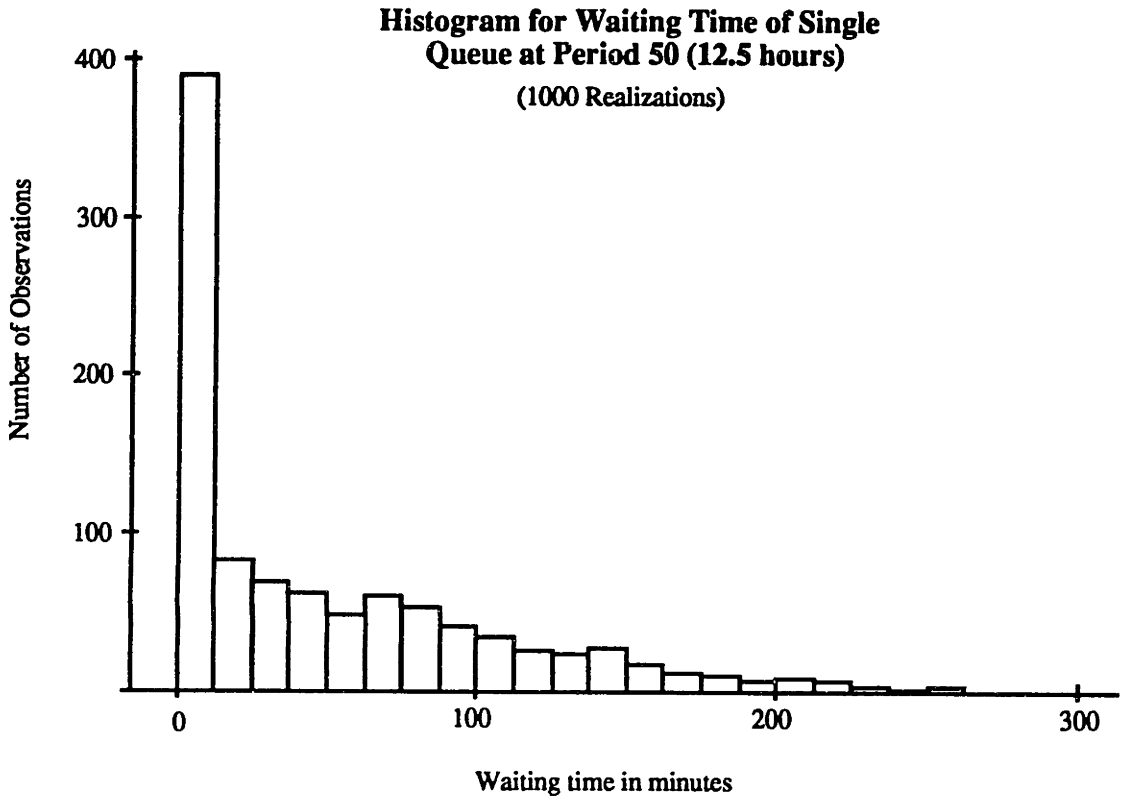


Figure 3.7: Histogram from simulated waiting times in a single queue

follow an approximately exponential distribution. This is confirmed in Figure 3.8, 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.[2] Plots such as this one suggest an approximate mixed distribution for the waiting times W_k:

$$\begin{aligned} \Pr\{W_k = w_{\min}(k)\} &= \delta \\ \Pr\{W_k \leq w \mid w > w_{\min}(k)\} &= 1 - e^{-\nu(w - w_{\min})} \end{aligned} \tag{3.23}$$

[2]If the exponential is correct, $\exp(-\nu W^{(n)})$ are realizations of the reverse cumulative distribution $\overline{F}(w)$ and should behave like the reversed order statistics of a $U[0,1]$ distribution.

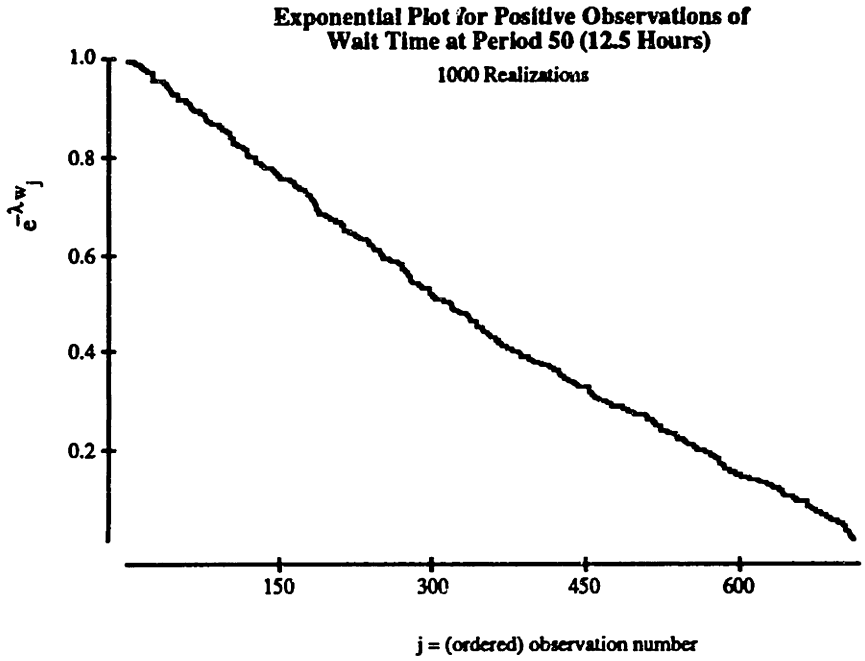


Figure 3.8: Test for exponential distribution of positive waiting time realizations

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 (subscripts omitted)

$$\begin{aligned} \delta w_{\min} + (1-\delta)\int_{w_{\min}}^{\infty} w\nu e^{-\nu(w-w_{\min})}\,dw &= E[W] \\ \delta\,(w_{\min})^2 + (1-\delta)\int_{w_{\min}}^{\infty} w^2\nu e^{-\nu(w-w_{\min})}\,dw &= E[W^2] \end{aligned} \tag{3.24}$$

This procedure will prove useful in the development of network algorithms in Chapter 5.

3.3.2 The Diffusion Approximation

The heavy loading of the system in our constant demand example is intentional, because under such conditions, a diffusion approximation of the transient queue behavior is possible. This method is applicable to a wide range of queueing systems

(including those where the arrival process and service process are correlated), but it does require a heavy traffic assumption.[3] It was developed in the 1970's [18,19] as a way of approximating a discrete-state queueing process by a continuous-state stochastic process more amenable to analysis. The summary presented here is taken from [12] and [17].

For a continuous time system on $t \geq 0$ define the arrival and service processes for a queue:

$$\begin{aligned} A(t) &\triangleq \text{cumulative number of arrivals up to time } t \\ D(t) &\triangleq \text{cumulative number of services up to time } t \\ Q(t) &\triangleq \text{no. in queue at time } t = [A(t) - D(t)]^+ \end{aligned}$$

Let $1/\lambda$ be the mean interarrival time and $1/\mu$ the mean service time. Define the traffic intensity $\rho = \lambda/\mu$. In the limit as $\rho \to 1$

$$Q(t) = A(t) - D(t).$$

Define

$$\begin{aligned} \Delta Q(t) &\triangleq Q(t+T) - Q(t) \\ &= A(t+T) - D(t+T) - [A(t) - D(t)] \\ &= \Delta A(t) - \Delta D(t). \end{aligned}$$

As $\rho \to 1$, for sufficiently large T the central limit theorem implies that $\Delta Q(t)$ is approximately normal:

$$\Delta Q(t) \sim \mathcal{N}[\beta T, \alpha T]. \tag{3.25}$$

The parameters β and α are called the drift and the diffusion coefficient and are defined by

$$\alpha \triangleq \lim_{T \to \infty} \frac{Var[Q(t+T) - Q(t)]}{T} \tag{3.26}$$

$$\beta \triangleq \lim_{T \to \infty} \frac{E[Q(t+T) - Q(t)]}{T}. \tag{3.27}$$

[3] Most applications of the diffusion approximation have also been conducted for a constant demand rate, as we do here. However, extension to time-varying demand is theoretically possible.

If the interarrival times and service times constitute independent series of independent random variables with variances σ_A^2, σ_S^2 respectively, it can be shown that

$$\begin{aligned} \alpha &= \lambda^3\sigma_A^2 + \mu^3\sigma_S^2 \\ \beta &= \lambda - \mu. \end{aligned}$$

Equation (3.25) implies that under heavy traffic, $Q(t)$ can be approximated by the continuous stochastic process $\{X(t), t \geq 0\}$ whose density function $f(x,t)$ obeys the Kolmogorov forward diffusion equation

$$\frac{\partial}{\partial t}f(x,t) - \beta\frac{\partial}{\partial x}f(x,t) + \frac{\alpha}{2}\frac{\partial^2}{\partial x^2}f(x,t) = 0 \tag{3.28}$$

Solution of (3.28) subject to initial and boundary conditions yields a solution for $f(x,t)$ or (equivalently) for the cumulative distribution function

$$F(x,t) \triangleq \Pr\{X(t) \leq x\}.$$

When the starting number in queue is x_0, the initial condition is

$$F(x,0) = \begin{cases} 0 & \text{if } x < x_0 \\ 1 & \text{if } x \geq x_0, \end{cases} \tag{3.29}$$

and the boundary condition is

$$F(0,t) = 0 \quad x_0 > 0, \quad t > 0 \tag{3.30}$$

which reflects the heavy traffic assumption that the queue is always non-empty.

Under conditions (3.29) and (3.30) the solution to (3.28) is

$$F(x,t) = \Phi\left(\frac{x - x_0 - \beta t}{\sqrt{\alpha t}}\right) - e^{2x\beta/\alpha}\Phi\left(\frac{-x - x_0 - \beta t}{\sqrt{\alpha t}}\right), \tag{3.31}$$

where Φ denotes the standardized cumulative normal distribution.

By differentiating (3.31), one obtains the density, from which it is possible to get $E[X(t)]$, an approximation of the expected queue length at time t, $E[Q(t)]$. Finally, if $\beta < 0$,

$$F(x) = \lim_{t\to\infty} F(x,t) = 1 - e^{-2x(-\beta)/\alpha} \tag{3.32}$$

and thus

$$\lim_{t\to\infty} E[X(t) \mid X(0) = x_0] = \frac{-\alpha}{2\beta}. \tag{3.33}$$

3.3.3 Adapting the Diffusion Approximation

The diffusion approximation as presented is not appropriate for the queueing system of Figure 3.4 because it assumes independent service times. In contrast, our system has periods of different *service capacities* which change according to a Markov chain. To apply the diffusion approximation, it is necessary to find the appropriate expressions for the drift and diffusion coefficient parameters β and α.

Consider successive aircraft periods indexed by k. Let ξ_k denote the number of services in period k. Under the heavy traffic assumption, ξ_k is a random variable taking one of the values $\mu_1, \ldots, \mu_S$. Let the steady state capacity probabilities be given as $\pi_1, \ldots, \pi_S$, so that the time average capacity is

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

Define the cumulative arrival and service processes by

$$\begin{aligned} A(N) &\triangleq \sum_{k=1}^{N} \lambda_k \equiv N\lambda \\ S(N) &\triangleq \sum_{k=1}^{N} \xi_k. \end{aligned}$$

Under heavy traffic assumptions,

$$Q(N) \triangleq Q_N = A(N) - S(N).$$

The drift and diffusion coefficient parameters for this process are given by

$$\begin{aligned} \beta &= \lim_{M\to\infty} \frac{E[Q(M) - Q(0)]}{M} \\ \alpha &= \lim_{M\to\infty} \frac{Var[Q(M) - Q(0)]}{M}. \end{aligned}$$

Assume that $Q(0) = A(0) = S(0) = 0$. Then it is immediate from the definitions that

$$\begin{aligned} \beta &= \lim_{N\to\infty} \frac{E[A(N)]}{N} - \lim_{N\to\infty} \frac{E[S(N)]}{N} \\ &= \lambda - \lim_{N\to\infty} \frac{E[S(N)]}{N}. \end{aligned} \tag{3.34}$$

Moreover, from the independence of the arrival and service processes and the fact that the arrival process is deterministic it follows that

$$\begin{aligned} \alpha &= \lim_{N\to\infty} \frac{Var[A(N)]}{N} + \lim_{N\to\infty} \frac{Var[S(N)]}{N} \\ &= \lim_{N\to\infty} \frac{Var[S(N)]}{N}. \end{aligned} \tag{3.35}$$

Thus the problem reduces to that of finding the values

$$\begin{aligned} \overline{S} &\triangleq \lim_{N\to\infty} \frac{E[S(N)]}{N} \\ \sigma^2(S) &\triangleq \lim_{N\to\infty} \frac{Var(S_N)}{N}. \end{aligned}$$

To do this we require a result concerning a central limit theorem for additive processes on a Markov chain due to Keilson and Wishart [21,22].

Theorem 3.10 *(Keilson and Wishart, 1963). Let $J(k)$ be a Markov chain with transition matrix $\mathcal{P} = \{p_{ij}\}$, and let $\xi(J(k))$ be a series of independent random variables with distributions determined by the state of the chain, $\xi_j \sim F_j(x)$. Define an additive process recursively by*

$$S(k+1) = S(k) + \xi(J(k)),$$

and define the matrices

$$\begin{aligned} \mathcal{B}(x) &\triangleq \{p_{ij}F_j(x)\}. \\ \mathcal{B}_n &\triangleq \int_{-\infty}^{\infty} x^n d\mathcal{B}(x) \end{aligned}$$

Suppose that $J(k)$ is ergodic and $S(k)$ is non-degenerate, in the sense that increments over regenerative cycles of J are not identically 0. Suppose also that the increment random variables $\{\xi\}$ all have finite second moments. Let

$$m = \lim_{k\to\infty} \frac{E[S(k)]}{k}$$

and

$$\sigma^2 = \lim_{k\to\infty} \frac{\mathrm{Var}[S(k)]}{k}.$$

Then

$$\frac{S(k) - km}{\sigma\sqrt{k}} \xrightarrow{d} N(0,1), \tag{3.36}$$

where $N(0,1)$ denotes the unit normal distribution. Moreover, the parameters m and σ^2 are given by

$$m = \pi^T \mathcal{B}_1 \mathbf{1} \tag{3.37}$$

$$\sigma^2 = \pi^T \mathcal{B}_2 \mathbf{1} - 3m^2 + 2\pi^T \mathcal{B}_1 \mathcal{Z} \mathcal{B}_1 \mathbf{1}, \tag{3.38}$$

where $\mathbf{1}$ denotes the vector of 1's and $\mathcal{Z}$ is the fundamental matrix of Markov chains,

$$\mathcal{Z} = \left[\mathcal{I} - \mathcal{P} + \mathbf{1}\pi^T\right]^{-1}.$$

The significance of the theorem in the present context is that it furnishes formulas for $\overline{S}$ and $\sigma^2(S)$ for the service process defined on the Markov chain. We have only to identify $\overline{S}$ with m and $\sigma^2(S)$ with σ^2 and apply equations (3.37) and (3.38). Note that in our case,

$$\mathcal{B}_1 = \{p_{ij}\mu_j\}$$

$$\mathcal{B}_2 = \left\{p_{ij}\mu_j^2\right\}.$$

Writing out (3.37) in full, we find that

$$\begin{aligned} \overline{S} &= \sum_j \sum_i \pi_i p_{ij} \mu_j \\ &= \sum_j \pi_j \mu_j \\ &= \overline{\mu}. \end{aligned}$$

Thus from (3.34) we may characterize the drift β as

$$\beta = \lambda - \overline{\mu}, \tag{3.39}$$

while the diffusion coefficient is

$$\alpha = \sigma^2. \tag{3.40}$$

3.3.4 Discussion

The preceding analysis indicates how a diffusion approximation may be given for a single queue with a constant arrival rate and a service process which varies according to a Markov chain. Recall the formula (3.31) for the queue length distribution. By finding the parameters β and α and using numerical integration, it is a straightforward matter to compute $E[X(t)]$, the expected value of the queue length approximation over time. This can be compared with $E[Q_k]$, the expected queue lengths calculated by the recursive algorithm. Figure 3.9 plots both sets of numbers over a 50-hour period for a heavily loaded system ($\rho \approx 0.95$). Also plotted is the asymptotic value of the expected queue length,

$$\lim_{t\to\infty} E[X(t)] = \frac{-\alpha}{2\beta} \approx 200.$$

The similarity of the curves suggests that the diffusion approximation captures the essential qualitative behavior of this queueing system. The approach to equilibrium is remarkably slow, with the limiting value of 200 not attained even after 50 hours (200 periods). While the true steady state value is not known, these observations suggest that the true queue length process is also very slow to converge.

In a system where the arrival rate is near constant and the traffic is heavy, transient behavior is described fairly adequately by a diffusion approximation, which may be computed in much less time than the recursion. For time-varying systems, one could carry out the approximation within periods of constant demand, using the previous period's final condition as an initial condition. A better method, however, would be to adapt the approximation to a time-varying arrival rate. Such an

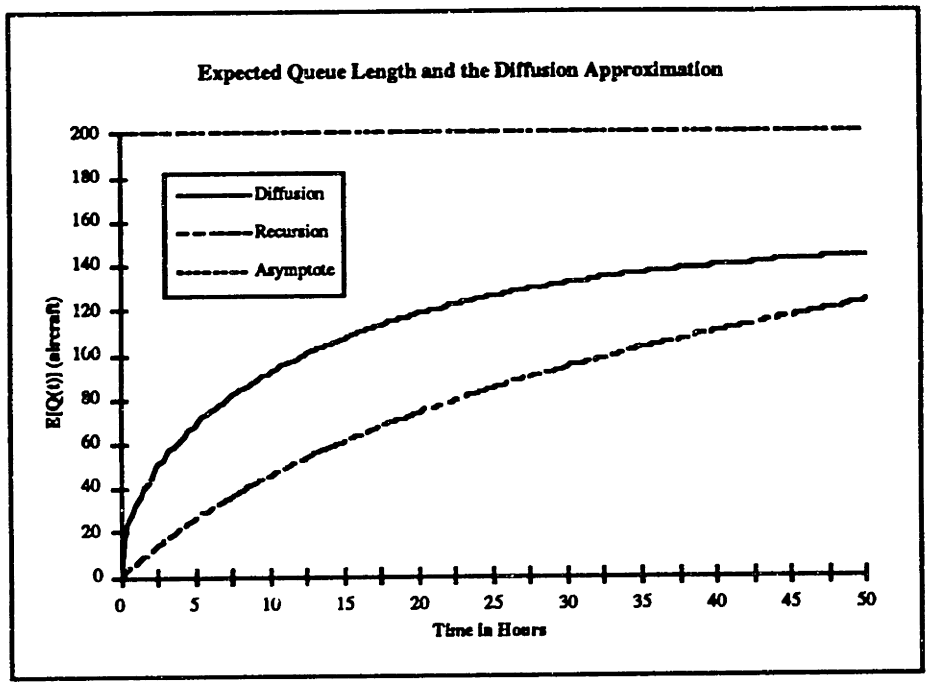


Figure 3.9: Comparison of expected queue length predicted by recursion and by the diffusion approximation

approach would be particularly useful in a network context, where speed becomes more critical.

Chapter 4

Dallas-Fort Worth: A Case Study

In this chapter we discuss an application of the one-hub recursive model of Chapter 3 to the case of Dallas-Fort Worth International Airport. Our intent is twofold. First, we hope to gain further insight into congestion at hub airports, and second, we hope to illustrate the usefulness of our congestion model in addressing important questions of policy.

The discussion is organized into four sections. In Section 4.1 we give necessary background on operations at DFW, focusing most attention on the service process for arriving aircraft. In Section 4.2 we discuss parameter estimation and assess the degree to which the data allow a simpler Markov (vs. semi-Markov) formulation. In Section 4.3 we present initial computational results and results of a limited validation using delay data obtained from the U.S. Department of Transportation (DOT). Finally, in Section 4.4 we present the main part of our results. Section 4.5 contains concluding remarks.

4.1 Operations at Dallas-Fort Worth

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

Traffic Type	Major Operators	Arrivals per Day
Air carrier	American, Delta	750
Air taxi	American Eagle, Atlantic Southeast	220
Military	—	less than 20
General	—	10-30

Table 4.1: Major demand sources at DFW. Source: Dallas-Fort Worth Airport Authority

terms of delay problems, with only the three New York area airports, San Francisco, and Chicago having significantly greater numbers of delays in 1989.[1] 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.

The Arrival Process

Arrival traffic at DFW falls into four categories, as illustrated in Table 4.1. The largest of these, air carrier traffic, consists of scheduled jet service and is dominated by the two hub carriers. The second largest category is air taxi service, which accounts for most propeller aircraft at Dallas. Much of this traffic is operated by the two commuter companies, American Eagle and Atlantic Southeast, which feed the jet service of their respective business partners, American and Delta. The remaining 30 or so landings at DFW per day consist of military aircraft and general aviation.

A typical daily demand schedule is illustrated in Figure 4.1. 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. The figure also includes the small numbers of military and general aviation aircraft which use DFW. Because these do not follow a regular schedule, we assume them to be spread uniformly through the major portion of the operating day (7 a.m. to 10 p.m.). The small number of such flights (20-30 per day) makes this assumption one of minor significance.

[1] *Winds of Change*, p.212.

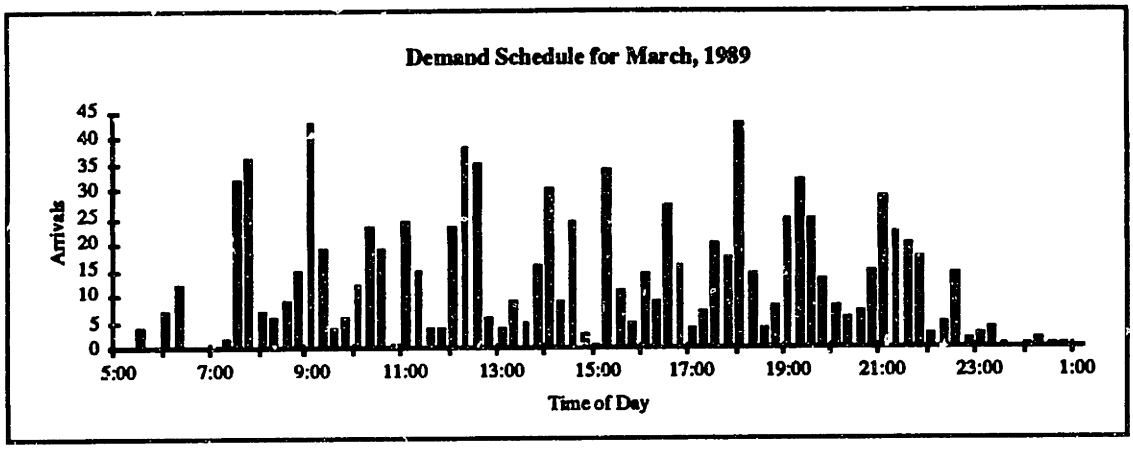


Figure 4.1: Arrival schedule at DFW for March 1989. Sources: DOT, OAG., and DFW Airport Authority

The Service Process

Consider next the service process for arriving aircraft at DFW. The recursive model of Chapter 3 requires that the user specify the parameters $\mu_1, \ldots, \mu_S$ and furnish a description of the underlying probabilistic structure. To do this, one would ideally like to have an historical record of available capacity. Unfortunately, no such record is available from any of the sources — the Federal Aviation Administration (F.A.A.), the Airport Authority, American Airlines — where one would expect to find it. Thus capacity specification requires a further examination of actual operations.

Figure 4.2 depicts the runway layout at DFW. There are four North-South runways[2] and two diagonal runways set off from these at an angle of 50°. During normal operations, one of each pair of runways is devoted to landings with the other used for takeoffs. Thus in favorable weather conditions DFW has three runways available to handle landing aircraft. However, in less favorable weather conditions, only two runways are available for landings, and capacity is correspondingly reduced.

Capacity under different configurations is given by "engineered performance

[2]The numbers marking the ends of each runway indicate (in tens of degrees) the compass direction taken by aircraft beginning their takeoff or landing at that end. There are four N-S runways, so to distinguish the second pair, the notation 17R/35L and 17L/35R is used.

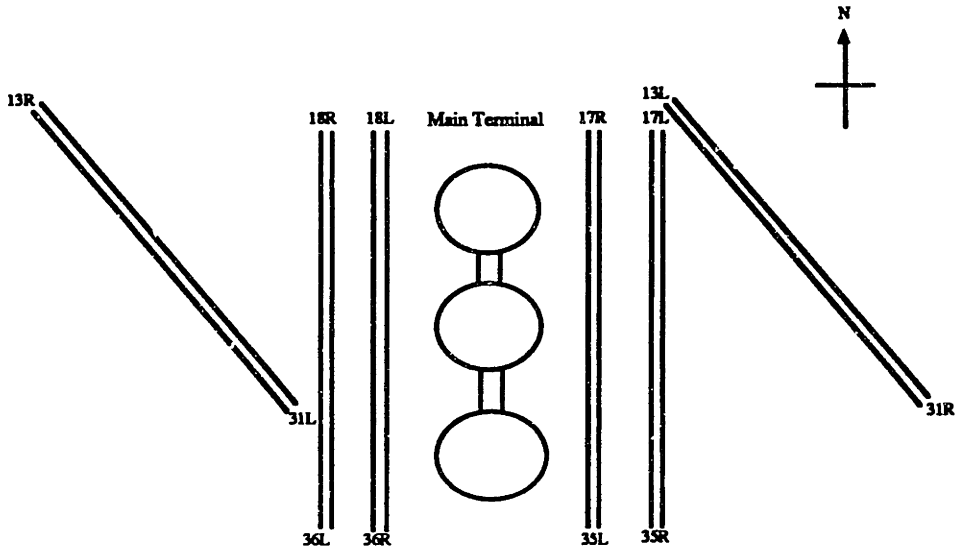


Figure 4.2: Map of runway system at Dallas-Fort Worth Airport.

Capacity under different configurations is given by "engineered performance standards" for the runways. These estimates are determined at the time of construction based upon known operating procedures under the different configurations. Essentially, they reflect the number of runways available and on how much separation is required between incoming aircraft.[3]

Federal regulations prohibit the use of a runway when the component of wind velocity perpendicular to the direction of travel exceeds 15 mph (for propeller-driven aircraft) or 20 mph (for jets). For this reason, runways are typically positioned so that the prevailing winds run parallel to them. At Dallas, prevailing winds tend to be North-South, and there are very few days when the four parallel runways cannot be used. Operations are more often shut down on the diagonal runways because these experience cross-winds more frequently. These runways also cannot be used during periods of low cloud ceiling and visibility, when the required separation between incoming aircraft increases. Ceiling and visibility are classified into the four states given in Figure 4.3: Visual Flight Rules I, Visual Flight Rules II, Instrument Flight

[3]A more detailed discussion of this process is given in [27].

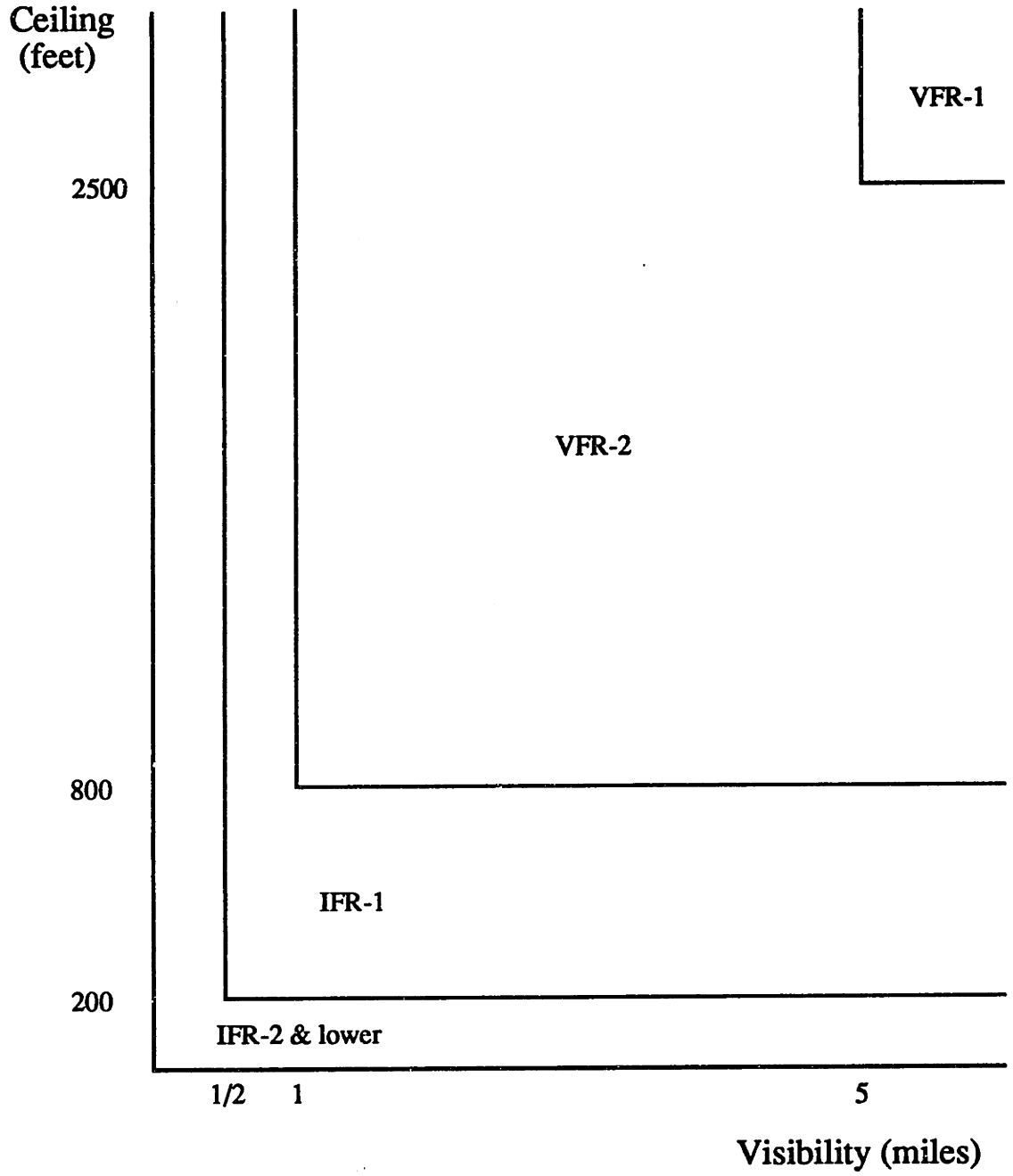


Figure 4.3: The four flight rules specifications

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

Rules I, and Instrument Flight Rules II & lower. In general, the higher the required separation, the lower the capacity.

The division of capacity into discrete states involves a degree of arbitrariness. Considering wind speed, wind direction, ceiling, and visibility together, we chose a total of six capacity states for DFW. Table 4.2 lists these six states together with the associated engineered performance standards (EPS) in aircraft per hour. As may be seen from the table, the capacity configurations range from the lowest state ('A') of 50 aircraft per hour up to the highest state ('F') of 95 aircraft per hour. There is a substantial difference between the two highest capacity states and all other states, due to the availability of the third runway.

Within the air transportation industry, EPS estimates are considered to be conservative for high-capacity configurations because under visual flight rules and good conditions, experienced pilots can operate safely with separations less than those assumed in setting the standards. On the other hand, the standards are not considered to be conservative for the lower-capacity states. To compensate, our implementation considers an ongoing study by UNISYS Corporation [13] which presents estimates of runway capacity per hour based on empirical observations made during peak periods. Preliminary results put the highest arrival capacity state at DFW in the range of 115 aircraft per hour, a substantial increase over the number 95 reported in the table. 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 con-

figuration estimates should remain essentially unchanged. We adopt these changes for the capacities in this study and note that the need for more accurate capacity estimation procedures seems obvious.

4.2 Estimation of Weather Change Parameters

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 4.2 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. There are considerable differences in average capacity from month to month. Figure 4.4 plots the average number of hours observed per month for three different capacities: 'A' (lowest IFR), 'D' (windy, VFR-1), and 'F' (still, VFR-1). State 'F' is dominant: its number of hours observed per month is greater than the others' by an order of magnitude.

Seasonal variability is evident in the data. Not surprisingly, lower visibility conditions tend to occur more in the winter; indeed, in summer, occurrences of this worst state are exceedingly rare. January, February, March, and April constitute the windiest portion of the year. Because of this seasonal variation, we chose a particular month (March) and based the parameter estimates on data for that month only. We chose March because its weather falls in between that of the low-capacity winter months and the high-capacity summer months. Configuration 'F' constitutes about 75% of March observations.

4.2.1 Estimation in the Markov Case

Suppose that the six capacity states ($i = 1, \ldots, 6$) follow a homogeneous Markov chain, with transitions occurring from one 15-minute period to the next according

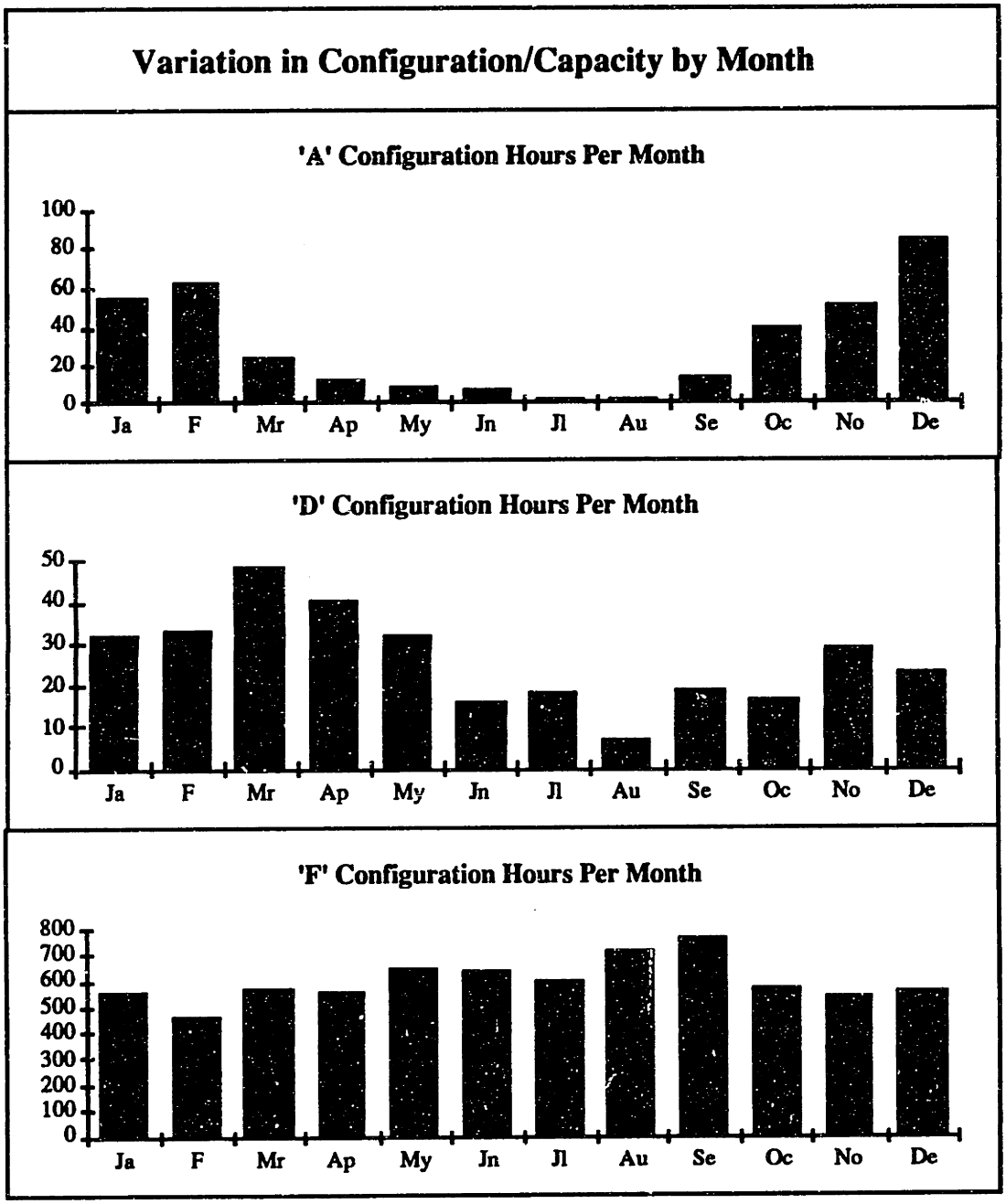


Figure 4.4: Capacity at DFW by month of year

to the transition matrix

$$\mathbf{P} = \{p_{ij}\}. \tag{4.1}$$

If observations were taken every fifteen minutes rather than every hour, then a sufficient statistic for estimating the transition probabilities would be the matrix $\mathbf{N} = \{n_{ij}\}$ of the number of transitions observed between states, and one could obtain the maximum likelihood estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}. \tag{4.2}$$

With hourly observations, we have only the matrix $\mathbf{N}' = \{n'_{ij}\}$ of hourly transitions. In theory, we could obtain maximum likelihood estimates from the likelihood function

$$\mathcal{L}(\mathbf{N}', \mathbf{P}) = \pi_{i_0} \prod_{i,j} \left(\sum_k \sum_l \sum_m p_{ik} p_{kl} p_{lm} p_{mj} \right)^{n'_{ij}}. \tag{4.3}$$

However, this estimation would require numerical methods, an amount of effort which seems unjustified given the arbitrariness of previous assumptions, such as $\Delta t = 15$ minutes. A simpler estimation procedure is suggested by the form of equation (4.2). The idea is to replace the unknown numbers n_{ij} by estimates $\tilde{n}_{ij}$ and then estimate the transition probabilities via

$$\hat{p}_{ij} = \tilde{n}_{ij} / \sum_j \tilde{n}_{ij}. \tag{4.4}$$

The estimates $\tilde{n}_{ij}$ are obtained from the n'_{ij} as

$$\tilde{n}_{ij} = n'_{ij} \qquad \text{for } i \neq j \tag{4.5}$$

$$\tilde{n}_{jj} = 4n'_{jj} + 1.5 \sum_i (n'_{ij} + n'_{ji}) \quad \text{for } j = 1, \ldots, 6. \tag{4.6}$$

The formula (4.6) has an intuitive explanation. Suppose there are $m+1$ consecutive hourly observations of state j preceded by some other state i and followed by some other state k. In symbols, this means $n'_{jj} = m$ (see Figure 4.5). Assume that if $n'_{jj} = m$, then the capacity has been in state j continuously between the first and

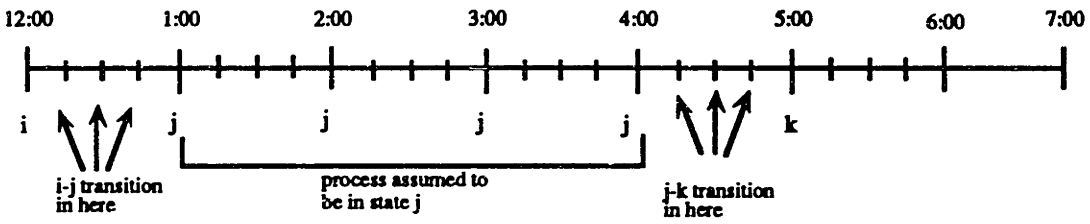


Figure 4.5: Schematic view of observation process for weather data

last of these observations.[4] Thus there are $4m$ self-transitions for 15-minute periods plus any just prior to the first hourly observation and just after the final one. The number of $j-j$ transitions may be thought of as a random variable

$$N_{jj} = 4n'_{jj} + X_a + X_b,$$

where X_b and X_a are random variables indicating the number of $j-j$ transitions associated with the hour just before the first of the $m+1$ observations and the hour after the last. We assume that, given that a transition took place at some point between hourly observations, it was equally likely to have taken place anywhere in that hour.[5] Thus

$$E[X_a] = E[X_b] = 1/4(0+1+2+3) = 1.5, \tag{4.7}$$

and the formula (4.6) follows.

4.2.2 Estimation for the Semi-Markov Model

Estimation for the semi-Markov model is more complicated than for the Markov model since the former requires two sets of parameters: the transition matrix $\mathcal{P}$ and

[4]This simplification neglects the possibility of multiple state changes between successive hourly observations. The result is to introduce a slight conservative bias in the estimates, in the sense that they reflect higher holding probabilities than is truly the case. However, the alternative is to compute all 4-period paths in the Markov chain and solve the likelihood equation numerically.

[5]Recall the earlier assumption of only one state change allowed between successive but different hourly observations. The justification for this new uniformity approximation lies in the memorylessness inherent in the Markov chain.

the holding time probabilities $\Pr\{T_i = m\}$. The matrix $\mathcal{P}$ has the natural estimator

$$\hat{P} = \begin{pmatrix} \hat{p}_{11} & \cdots & \hat{p}_{16} \\ \vdots & \ddots & \vdots \\ \hat{p}_{61} & \cdots & \hat{p}_{66} \end{pmatrix}. \tag{4.8}$$

where the $\hat{p}_{ij}$ are given by

$$\hat{p}_{ij} = \begin{cases} n'_{ij} / \sum_{j \neq i} n'_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{4.9}$$

Notice that the difficulty with estimating self-transitions is removed.

The holding time probabilities can be calculated from the observed hour-based histograms, but there is again the problem of inferring parameters for 15-minute periods from hourly observations. To overcome this, we elected to assign equal probability masses to all lengths in between the hours. For example, if the observed probability masses for 3 hours and 4 hours are p_3 and p_4, then the inferred probability masses for 3 hours, 3 hours 15 minutes, 3 hours 30 minutes, and 3 hours 45 minutes are $(p_3 + p_4)/8$. Similarly, lengths between 4 and 5 hours have the estimates $(p_4 + p_5)/8$ and so on. The fact that the inferred probabilities sum to 1 is easily checked.

The implicit uniformity assumption here has less justification than in the Markov model, but without some hypothesis about the underlying distribution, there can be no further structure added.

4.2.3 Evaluation of the Markov Model

Recall from the discussion of Chapter 3 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

$$N(i) \triangleq \sum_{m \geq 1} N(i,m). \tag{4.10}$$

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 4.6, 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 five of the six cases, geometric distributions appear to fit the data fairly well, although in every case there is greater actual probability mass at 1 hour than the model predicts. The poorest fit occurs with state 'F'. In this case, actual observations of extended periods of good weather force the geometric model to give a rather flat distribution, which fails to capture the behavior at low values. With all states, the dropoff in the early part of the distribution is greater than the estimated geometric rate.

A formal way to test the fit is to perform a χ^2 test of the null hypothesis that the values $N(i,m)$ are distributed according to the Markov model. As Table 4.3 indicates, the results are not favorable — the computed statistics in all cases fall well out on the tail of the χ^2 distribution implied by the null hypothesis.

But do these results necessarily lead to a rejection of the Markov model? The answer is not clear. The Markov model, like all models, is an abstraction of reality,

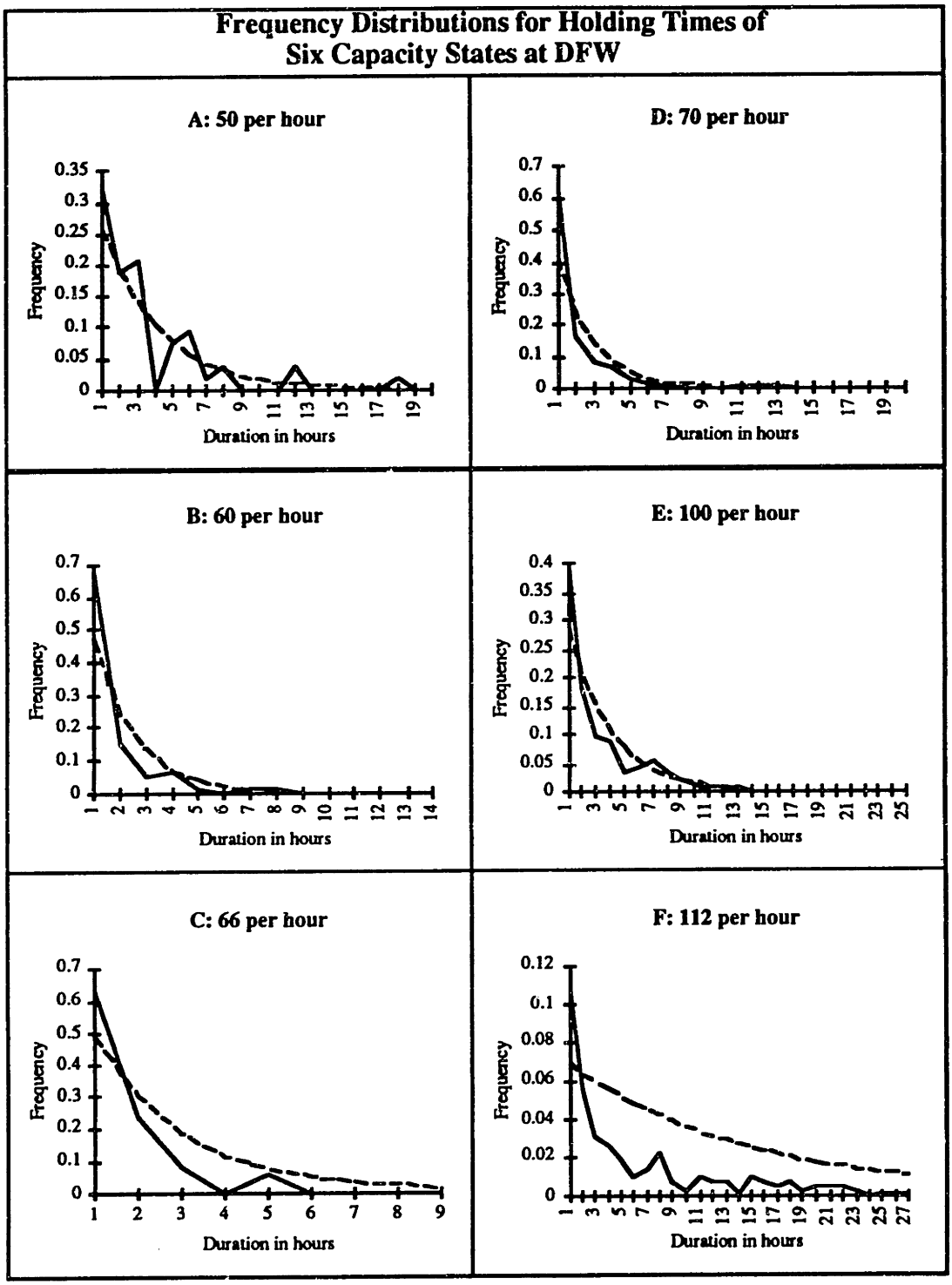


Figure 4.6: Examining goodness of fit for the Markov model. The solid lines indicate the observed frequencies for run lengths, while the dashed lines indicate the expected frequencies under a Markov chain model.

State	%tile from χ^2 test
A	0.80
B	0.985
C	0.995
D	0.995+
E	0.975
F	0.995+

Table 4.3: Results of χ^2 test for the six holding time distributions

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 4.4: Predicted and actual occupancy probabilities at DFW

and we should not necessarily expect it to do well in formal statistical tests such as this one. Moreover, while the holding times do not conform exactly to the data, the predicted state occupancy probabilities (i.e. the numbers π_i) are *extremely* close to the time-fractions observed in the data (i.e. the numbers $N(i)/\sum_i N(i)$ — see Table 4.4). Apparently, despite the fact that the holding times in each capacity state are not strictly geometric, the overall time fractions for each state are well predicted by a Markov chain. Considering the fact that the Markov model offers substantial computational savings, we are reluctant to reject it solely on the basis of the χ^2 test. A further verdict awaits test runs with actual traffic data.

4.3 Model Validation

Validation of the recursive queueing model presents considerable difficulties, because the data necessary to conduct rigorous tests are not readily available. The most comprehensive data for air traffic delay in the U.S. are the On Time Arrival Statistics

(OTAS) which airlines must report to the Department of Transportation. On a monthly basis the DOT receives this information for every flight performed by the major domestic jet operators, including scheduled departure time, actual departure time, scheduled arrival time, and actual arrival time. Unfortunately, this information is not sufficient for conducting precise tests for at least two reasons. First of all, the schedules against which the DOT measures delays are not reliable, because carriers have responded to the risk of poor on-time statistics by including slack in scheduled flight times. However, this difficulty can be resolved to some degree by normalizing flight times according to some standard. A much more serious difficulty with the data is that they reflect *total* aircraft delays rather than just queueing delays at the destination. These include:

1. *Departure delays.* Late arrivals at DFW coinciding with late departures from the preceding airport may be due to ground holds (DFW congestion) or to departure congestion. The ambiguity is not easily resolved.

2. *Travel time.* Delays calculated from the DOT data include travel delays caused by elements unrelated to congestion (e.g. head winds).

3. *Gate delays.* DOT data include delays caused by lack of available gate space for arriving flights. While this kind of delay clearly falls within the category of congestion at the destination, it is not included in our queueing model.

To summarize, the DOT data reflect *total delays*, including *delays carried over from earlier periods*, while our queueing model is concerned solely with *waiting times caused by runway congestion upon landing.* This discrepancy severely hampers efforts at validation. Indeed, the only real 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. Such an exercise is beyond the scope of the work undertaken here. Of course, a full-fledged implementation demands such procedures. For

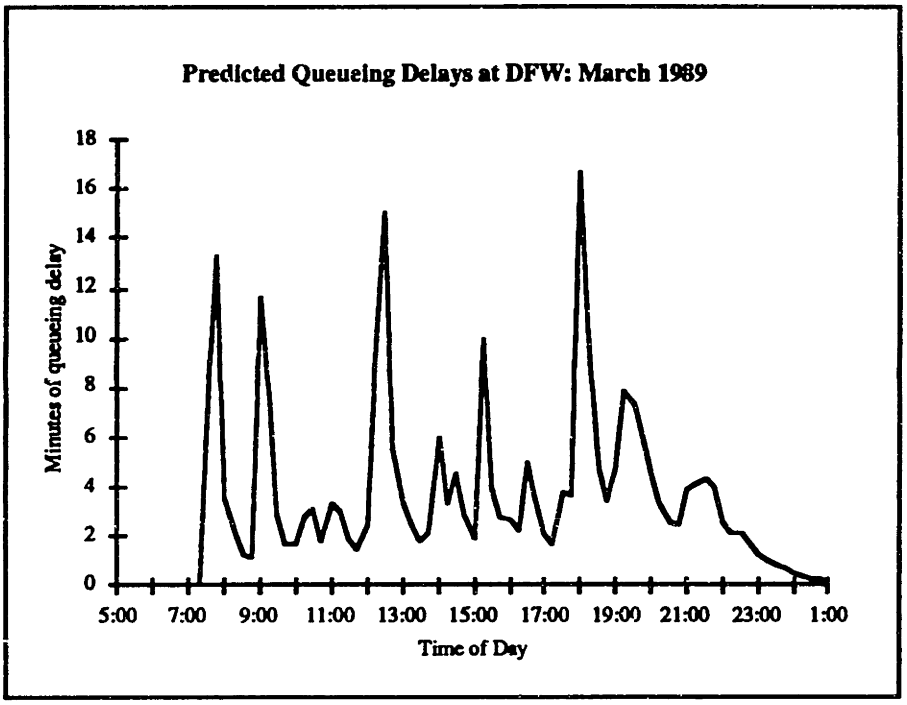


Figure 4.7: Expected waiting times at DFW based on March weather and 1989 traffic

the purposes of this study, subjecting the findings to some informal data analyses, keeping the above remarks in mind, must suffice.

Consider first the predictions of the queueing model. Figure 4.7 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 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 landing delay on average reaches 15 minutes, a good

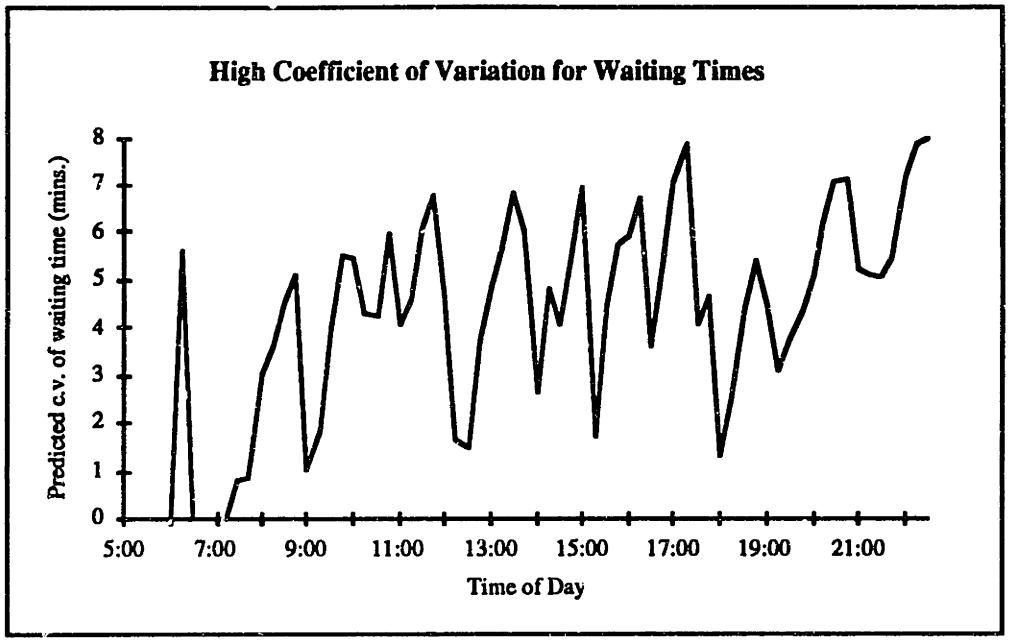


Figure 4.8: Coefficient of variation for delays at DFW under Markov model

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.

Waiting time variance is high. Figure 4.8 plots the predicted (unconditional) coefficients of variation in the waiting times

$$C_w(k) = \sqrt{\frac{Var(W_k)}{E(W_k)^2}}.$$

As may be seen, these values are substantially greater than unity, a reflection of the possibility of widely varying sample paths for the process, including the possibility of extreme tail values (extended periods of low capacity).[6]

[6]Compare the situation where capacity remains high throughout the day to that where it remains low. In the former case, there will be delays only at the busiest times, while in the latter case, capacity is inadequate for all but the lightest demand periods, and delay may reach into the range of several hours.

In order to conduct a validation test, we examined the DOT statistics for March, 1989. The relevant data are

$$\begin{aligned}
\text{DTC} &\triangleq \text{Scheduled Departure Time} \\
\text{DTA} &\triangleq \text{Actual Departure Time} \\
\text{DD} &\triangleq \text{Departure Delay} = \text{DTA} - \text{DTC} \\
\text{ATC} &\triangleq \text{Scheduled Arrival Time} \\
\text{ATA} &\triangleq \text{Actual Arrival Time} \\
\text{FTC} &\triangleq \text{Scheduled Flight Time} = \text{ATC} - \text{DTC} \\
\text{FTA} &\triangleq \text{Actual Flight Time} = \text{ATA} - \text{DTA}.
\end{aligned}$$

To correct for possible inconsistencies in scheduled flight lengths, for each origin we calculated a single average scheduled flight length (AFTC). Then for each flight i, we determined the total delay upon arrival as

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

Note that this statistic includes all possible flight delays, not only those due to landing congestion. To correct for outliers, we grouped all 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. In so doing, we hoped to omit observations reflecting long delays due to reasons other than congestion. We then ordered the remaining observations by scheduled arrival time, grouped them in 15-minute intervals (recall $\Delta t = 15$), and calculated means. These average *total* delays per 15-minute period are plotted in Figure 4.9 together with the average landing congestion delays predicted by the Markov model.

The results are somewhat ambiguous and point to the difficulties cited above. 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. On the other hand, the differences are larger at some times of the day

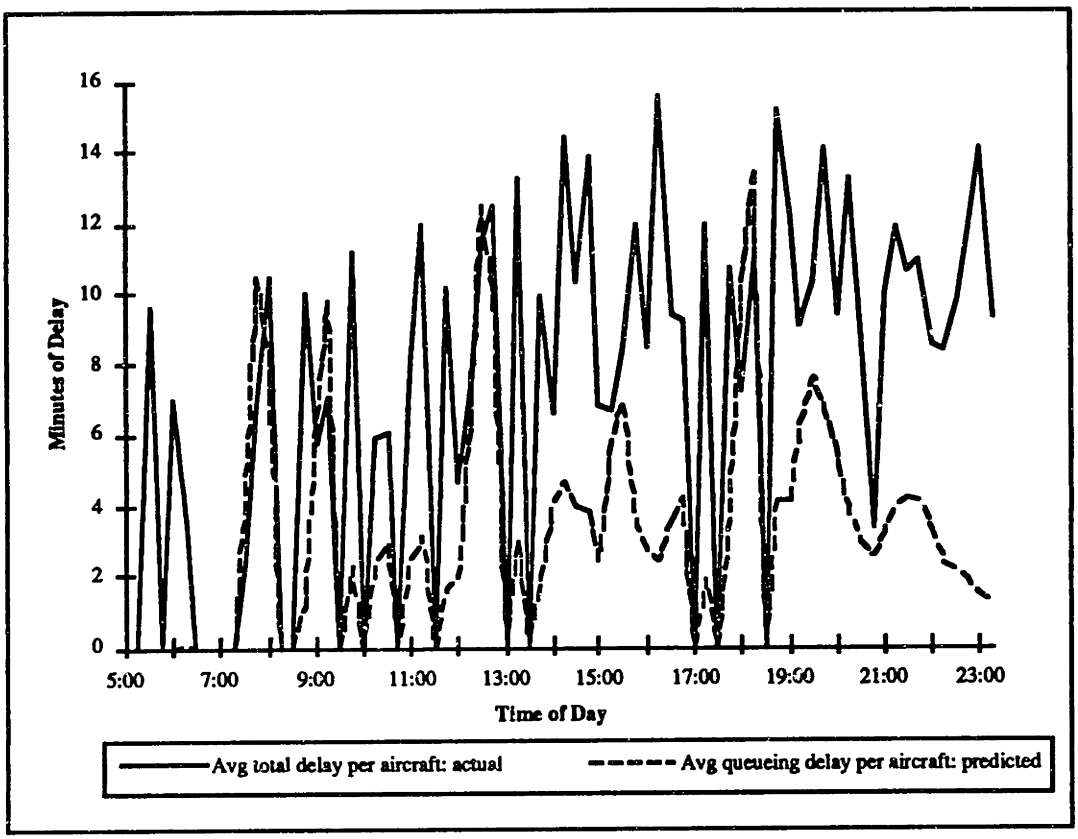


Figure 4.9: Predicted waiting times at DFW (from queueing model) compared with average total aircraft delays from adjusted DOT statistics

than at others, and several peaks exist where none are predicted. Travel time delays (e.g. head winds) seem the only possible explanation in some of these cases — for example, the earliest Delta and American banks of the day (around 5:30 a.m.) show significant delays, but the level of traffic at that time is nowhere near a level which would suggest significant congestion. A more likely explanation lies 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.[7]

Mid-morning and mid-afternoon discrepancies are more troubling. The banks at the latter of these times again correspond to arrivals from the west coast. Large actual total delays in the late evening cannot really be attributed to congestion at

[7]Airlines are more likely to hold flights at these times of day as a service for late passengers.

DFW, since the traffic at those times is very low. High lateness statistics here may be attributable to the fact that the originating departures are the last flights of the day (see the previous footnote), or more likely the fact that they reflect *delays carried over from earlier portions of the day.*

To give some idea of the magnitude of the differences between the two curves, we define the 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 Figure 4.9, this standard error is

$$s = 6.7 \text{ minutes},$$

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. These numbers indicate a fairly large difference between the two curves, with the major discrepancies coming in mid-afternoon and late evening.

On balance, these validation results are a better indication of the shortcomings in the data than of the accuracy of our queueing model. In the absence of a fully controlled validation experiment, however, 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.

4.4 Results and Discussion

In this section we explore some of the implications of the model's results at DFW. The following set of questions will guide the discussion:

- Are the results for the Markov and semi-Markov models appreciably different?
- Are the results of these stochastic models appreciably different from those obtained from a purely deterministic analysis?
- How do the correlations in service times implied by the model affect predicted delay?
- What does the model predict about how American and Delta affect one another at DFW in terms of congestion?
- What is the effect of schedule peaking on predicted delay?
- What are the advantages and disadvantages of traffic smoothing at DFW?

Markov vs. Semi-Markov Model

Figure 4.10 plots mean waiting times (averaged over initial conditions) for both the Markov and semi-Markov 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. This observation reinforces our earlier remarks about the limited suitability of strict hypothesis testing in this context. Apparently, the differences between the Markov and semi-Markov models constitute only second order effects at DFW. 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

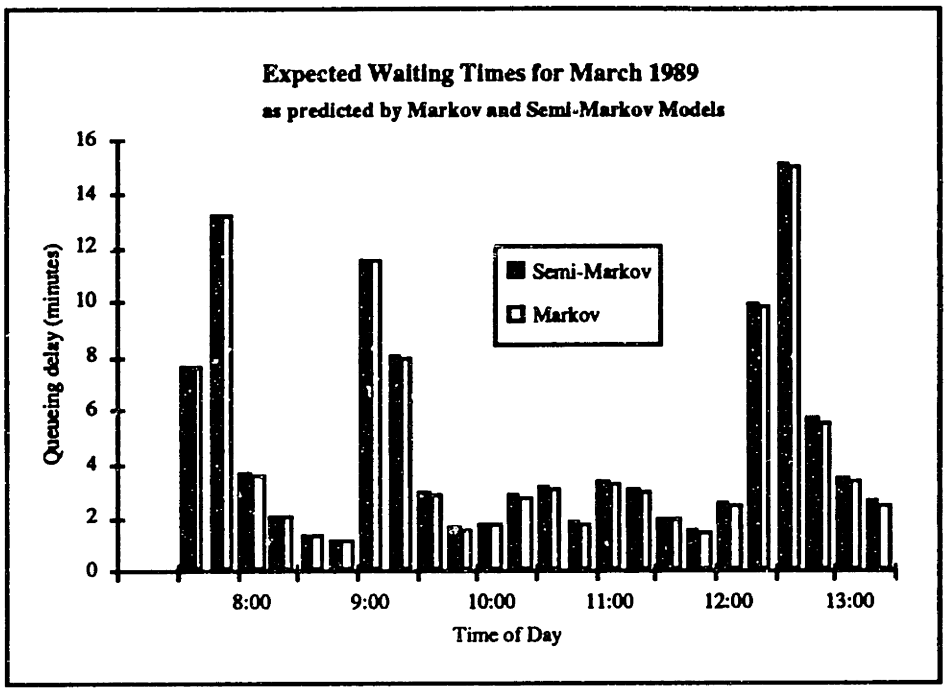


Figure 4.10: Comparison of predictions of expected waiting times at DFW under the Markov and semi-Markov models

An examination of the profiles predicted by the Markov and semi-Markov models suggests 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 our stochastic model compare with a deterministic analysis like that of Chapter 2? As an answer, consider Figure 4.11. 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

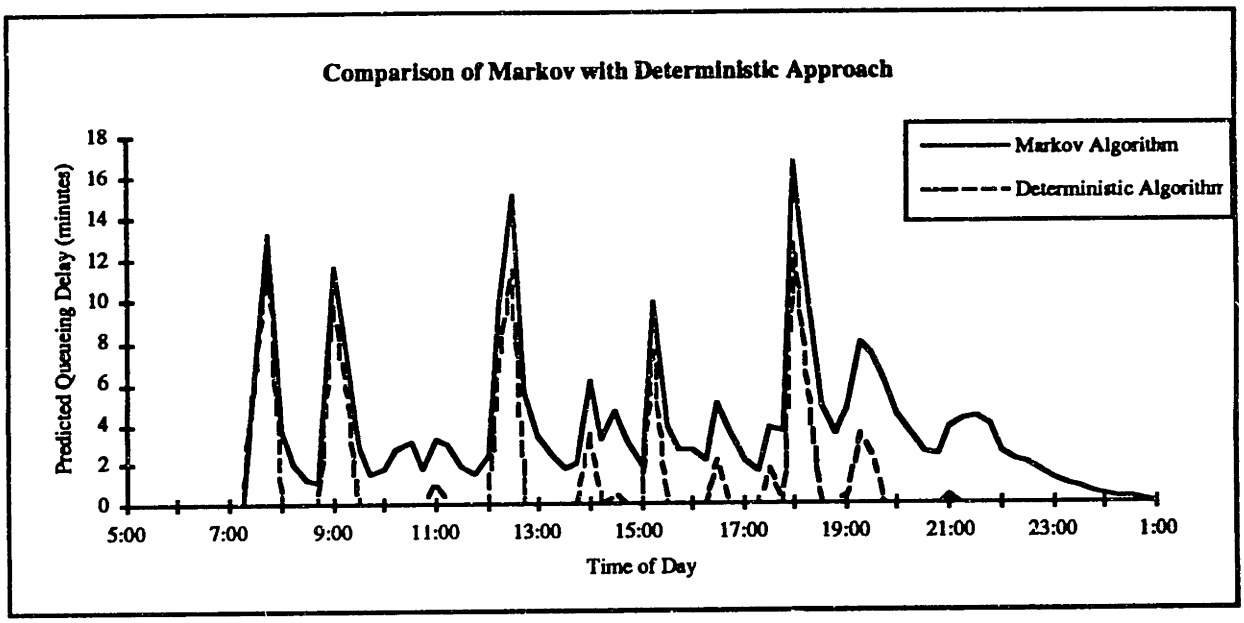


Figure 4.11: Comparison of predictions of expected waiting times at DFW under Markov and deterministic models

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 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

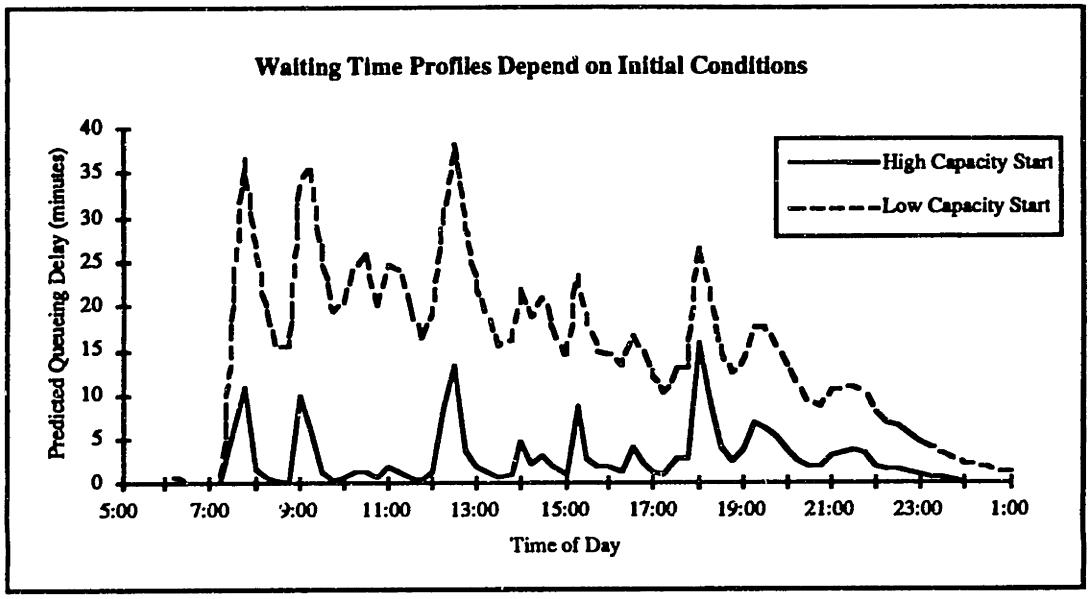


Figure 4.12: Capacity correlation means that initial conditions are important in determining expected waiting times.

in turn implies that mean queue lengths and waiting times will look quite different conditional on different starting states. Figure 4.12 plots two waiting time profiles based upon the starting states 'A' (lowest capacity) and 'F' (highest capacity). The difference is striking, with waiting times in the former case higher by an approximate factor of 3 throughout the day. 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 of a given state i in any period is equal to the steady state probability π_i. The effect of this new model is to eliminate correlation from period to period.[8] This change should reduce predicted mean waiting times, a fact which is confirmed by Figure 4.13. Note that the Markov model has only slightly higher estimates than the independent model for *peak periods* — the deterministic effect once again.

[8] However, within a given interval, all service times are in some sense still perfectly correlated; that is, when capacity enters state i, all customers served in that interval still have identical service times.

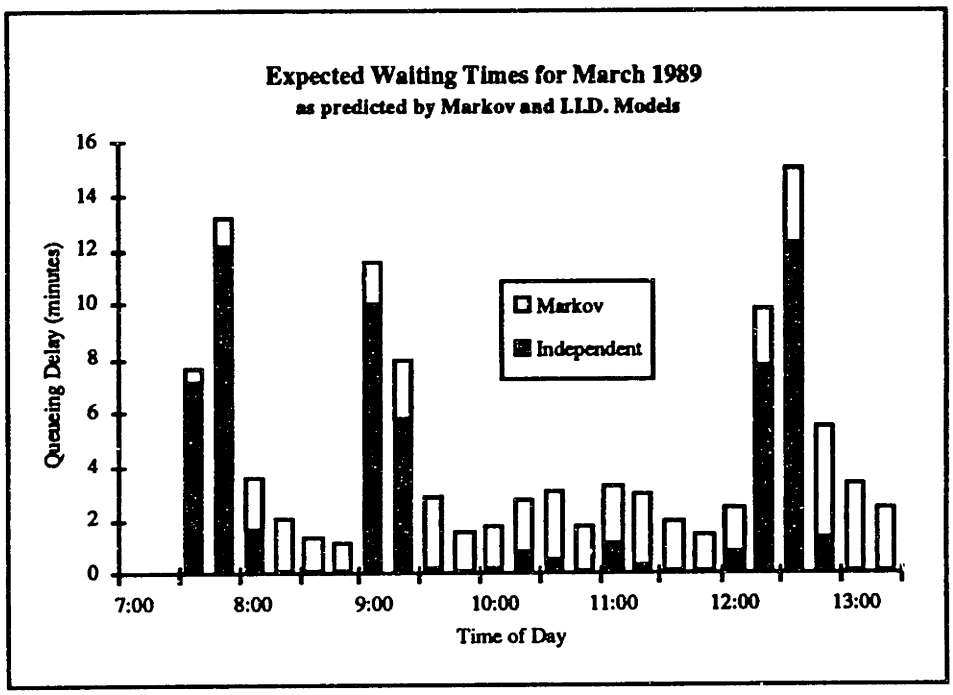


Figure 4.13: Comparing Markov and i.i.d. models illustrates the effects of correlations in capacity from period to period.

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) we have already seen that 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 4.14. Here, we

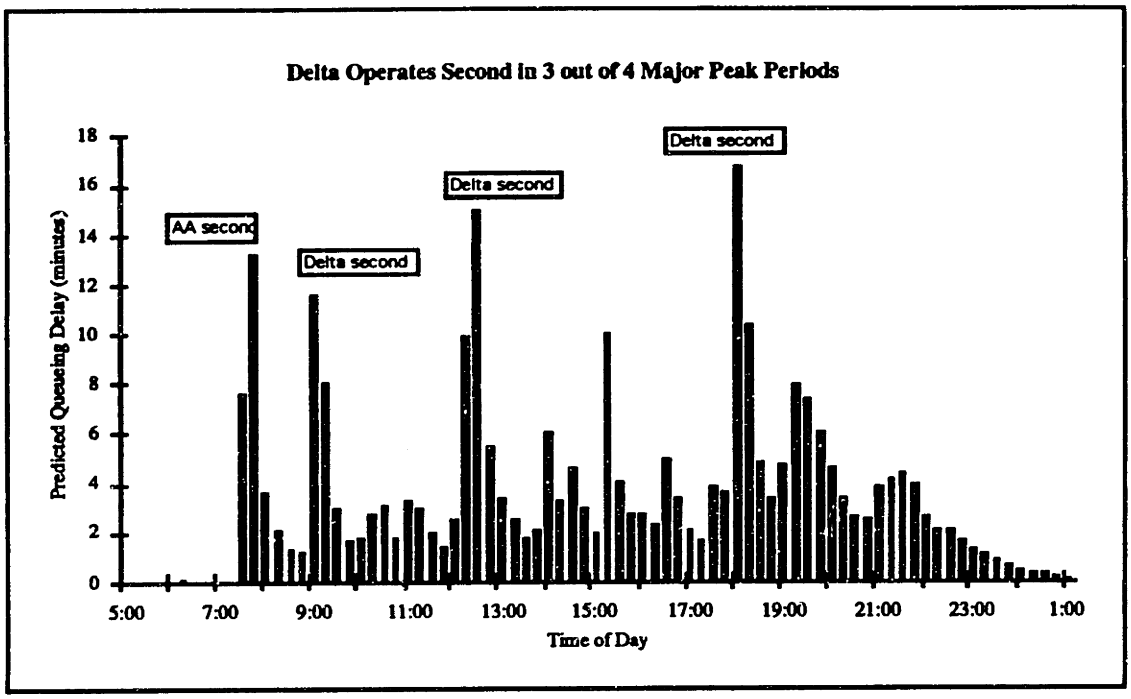


Figure 4.14: The four major double banks at DFW, labeled with the 2nd scheduled carrier in each case. Although both major carriers at DFW are affected by delays, Delta may bear a higher risk of waiting since its peaks are mostly scheduled right after American's.

have labeled the four highest delay peaks where the two carriers have arrival banks in close proximity. In each case, the label indicates 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.[9]

As an experiment with 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

[9]To improve the situation, Delta could of course alter its schedule, but there are other factors which work against this change. For example, the 6 p.m. peak involves numerous aircraft in the midst of a west-to-east bank. Delaying the departure of these aircraft might have significant costs in marketing, since such action would delay the eventual arrival times on the east coast, which are already quite late — about 10 p.m.

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

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).[10]

Table 4.5 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 tend to increase American's queueing delays. In the case where the two carriers' banks actually overlap significantly (bank 3), Delta shows higher 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

[10]Note that in this case, because we are comparing carriers rather than evaluating absolute estimates, the difficulties discussed earlier are less significant. That is, here we are concerned only with whether or not a difference exists rather than with the absolute numbers. It is important, however, that we correct for schedule stretching, since it is possible that one carrier practices this more than the other.

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.[11] 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 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 4.15, 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.[12] The $L = 20$ policy goes much further, introducing excess capacity approximately 85% of the time at Dallas. Appropriately, we term this level of smoothing "severe."

Figure 4.16 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 4.6. In moving from no smoothing to severe smoothing, there is a reduction in weighted average delay of

[11] For a more thorough discussion, see *Winds of Change*, Chapter 6.

[12] It is interesting to note that at certain times of day, scheduled demand actually exceeds the airport's highest capacity of 112 per hour (28 per 15-minute period). This phenomenon persists despite the reality that delay for these aircraft is a virtual certainty. Apparently, carriers have decided that the market benefits of serving the traffic at these peak times outweigh the costs of the resultant delays.

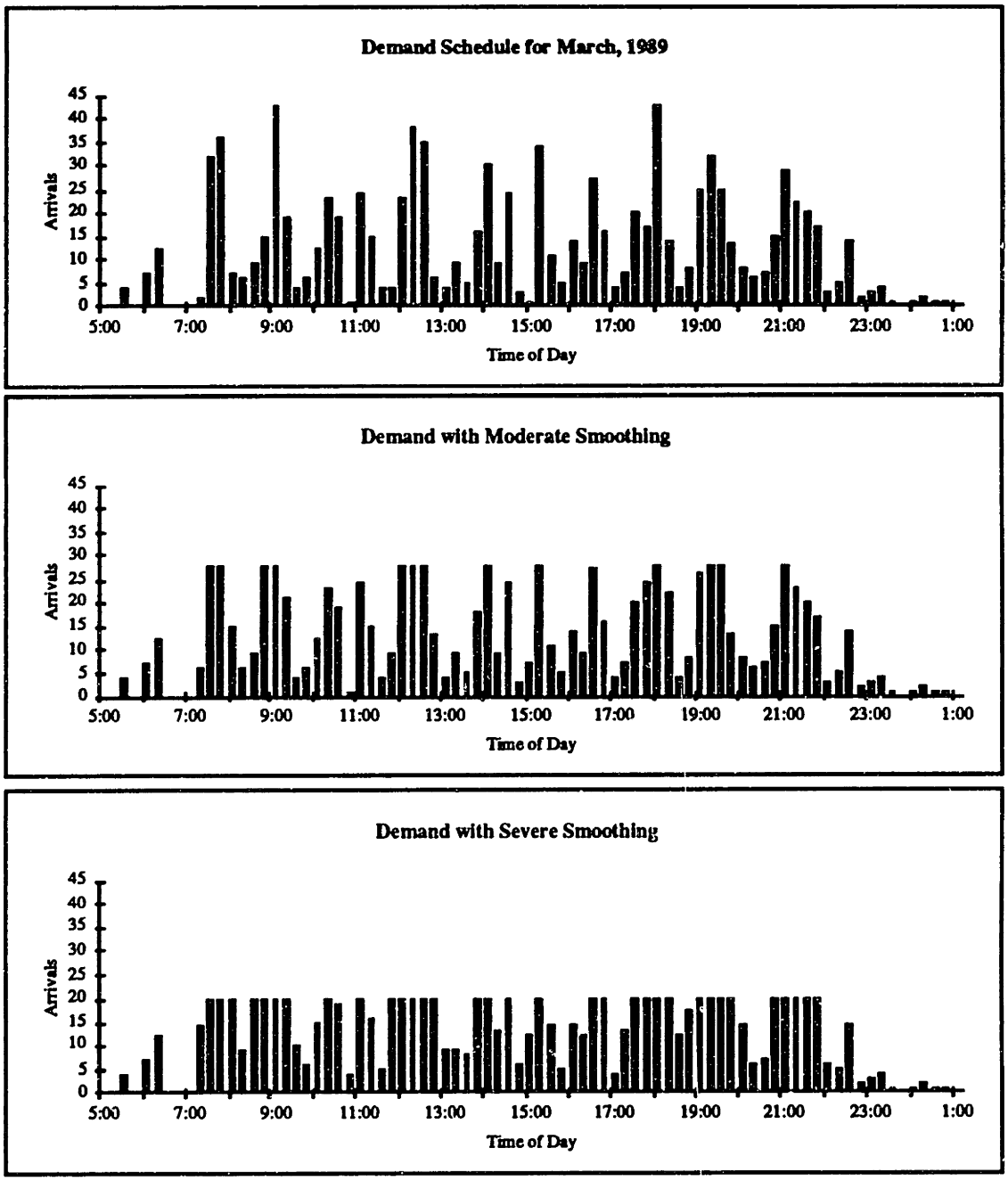


Figure 4.15: Alternative degrees of smoothing for DFW traffic

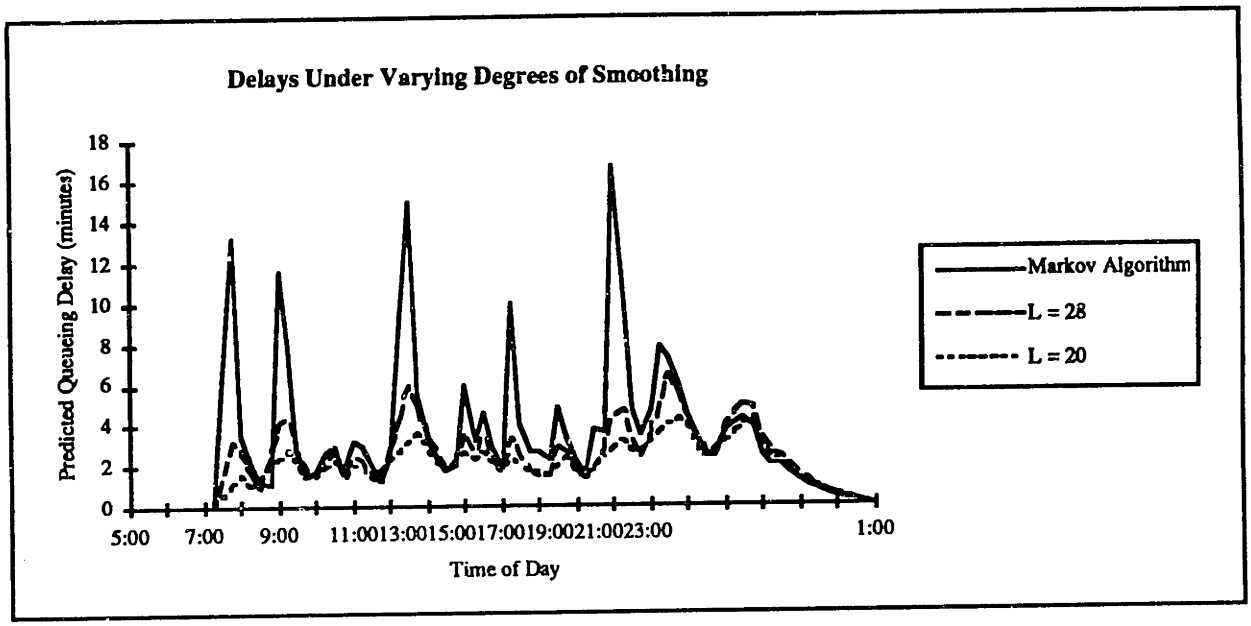


Figure 4.16: Predicted effects of traffic smoothing on waiting times

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

Table 4.6: Costs and benefits of smoothing policies

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 smoothing; reduction beyond this level of smoothing is relatively modest.* In other words, diminishing returns exist: once the schedule is smoothed to the point of "rationalization," most of the delay benefits have been realized, and gains from further smoothing are not as great. This result is consistent with the general principle in queueing theory that the greatest improvement in performance is obtained near $\rho = 1$.

The cost of the smoothing policies is difficult to assess. Some flights are shifted from the major banks, resulting in longer complex times. Table 4.6 lists the percent-

ages of flights shifted in 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.

But is any policy better than no policy? Newell [27] has remarked that delays themselves eventually work to police carriers operating at congested airports. If this is the case, active demand smoothing policies might well be viewed as unnecessary. Final evaluation must assess the likely response of carriers, including the possibility of their utilizing other hubs. Congestion models such as this one have a clear role to play, but the technical issue of congestion is only one part of the story.

4.5 Concluding Remarks

In this chapter we have attempted to demonstrate how the theoretical model developed in Chapter 3 can be implemented and used. In so doing, we have tried to indicate the necessity of careful attention to operational detail, especially in light of less-than-ideal conditions for estimating parameters. Unfortunately, available data are inadequate for conducting a thoroughly controlled validation procedure.

Analyses based on the model indicate a number of interesting features of this queueing system. First of all, as we have emphasized, the system exhibits a large amount of variability due to the great disparity between alternative sample paths. This high variance is reflected in the wide differences observed under different initial conditions. These differences reflect in turn the serial correlation inherent in the high self-transition probabilities estimated from the data. Our observations reinforce the idea of the necessity of transient analysis.

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, improved scheduling on the part of Delta (allowing itself greater slack at those times of day where it has

major banks near those of American) could improve performance. 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.

Chapter 5

Congestion Models for Networks

The natural extension of the work of the previous chapters is to consider the problem of congestion in a network. While isolated hub models are of obvious interest, airports operate in the environment of other airports, and congestion delay at one location affects performance at others.

This interaction between airports is particularly important for hubs in a hub-and-spoke network. Because of dependencies created by the schedule, performances at the hubs may be highly interconnected. For example, when American Airlines experiences a low capacity day at Chicago, the effects might be felt in Dallas, even if Dallas has no significant congestion problem. In fact, concerns about such an effect led American at one time to consider a policy of *isolating* the Chicago hub, requiring all or most departures from O'Hare to return to O'Hare rather than visit another hub in the system [10]. Although the strategy was never enacted, the mere fact that it was considered underlines the importance of the network problem.

In this chapter we will show how the methods of Chapter 3 may be extended to apply to networks of airports. Our goal is once again to develop models which improve our understanding of the system. More generally, our work constitutes a new approach to addressing a class of transient queueing network problems. These problems are of great interest in the operations research community because they

are both relevant in practice and difficult to solve.

The chapter is organized as follows. In Section 5.1 we describe the general queueing network context into which this airline problem falls and outline two decomposition approaches which exploit the recursive method for the single airport introduced in Chapter 3. Section 5.1.1 describes the first of these, in which downstream arrivals are adjusted according to expected upstream waiting times, while Section 5.1.2 describes a more involved approach which uses second moment information about delays to give a stochastic description of downstream arrival rates. In Section 5.2 we test the two methods against a simple simulation procedure on a 2-hub network. Our results indicate that the approximations inherent in the models work well for moderate traffic but tend to underestimate the spreading of demand which takes place in heavy traffic situations. These shortcomings suggest that further improvements would be welcome, particularly in the second method, which allows for more flexibility. We provide concluding remarks in Section 5.3.

5.1 Queueing Approaches for Networks

The network problem represents a significant increase in complexity over the problem of the single hub airport considered in Chapter 3. Recall that in the approach taken there, a known schedule of arrivals was specified as an input, and the outputs were the queue length and waiting time moments. In a network context, matters are complicated by the fact that delays at one airport alter the downstream arrival pattern. Moreover, an airline network is a *multi-class* queueing system, with the different classes being the individual aircraft. This is due to the fact that every aircraft has a unique *itinerary* specifying the details of its passage through the system. Thus our overall problem is one of describing the transient behavior of a multi-class queueing network with auto-correlated service rates at each node. This high degree of complexity suggests that approximation methods are necessary.

The following two subsections present two such approximation methods based on

decomposition, where we apply the recursive method to each airport in the network and use the resulting estimates of congestion delay to update itineraries and adjust arrival rates.

5.1.1 A Decomposition Approach Based on Expected Waiting Times

A simple decomposition approach is based on the following idea. Suppose that at the start of the day, one knows the schedules for all aircraft operating in the network. Under the assumption that delays are zero at the outset of the day, the schedule for the initial period of the day is fixed. Hence the first-period demands are fixed, and mean queue lengths and waiting times for each airport during this period may be determined by applying the one-hub recursive algorithm separately to each airport. The resulting expected waiting times for period 1 are estimates of the delay encountered by all aircraft scheduled to land in this period. Taking into account the slack which these aircraft have in their schedules and updating future arrival streams accordingly, one then fixes demand for the next period, calculates the resulting new expected waiting times, and so forth. We refer to this simple decomposition approach as "Decomposition Algorithm 1." Details are as follows.

Consider a network of airports $i = 1, 2, \ldots, N$. On this network let there be a set $\mathcal{A}$ of aircraft numbered $v = 1, 2, \ldots, V$. Divide the operating day into periods of length Δt, numbered as $k = 1, 2, \ldots, K$. Each aircraft v has an itinerary

$$\mathcal{I}(v) \triangleq \{(i_m^v, t_m^v, s_m^v)\} \quad m = 1, 2, \ldots$$

where

$i_m^v \triangleq m$th stop on itinerary of aircraft v

$t_m^v \triangleq$ scheduled arrival time at mth stop for aircraft v

$s_m^v \triangleq$ slack time between stops $m-1$ and m for aircraft v.

Let d^v represent the current cumulative delay for aircraft v — i.e. as one traces aircraft v through its itinerary, d^v represents the current amount by which it is

behind schedule. Further define the terms

$\mathcal{A}(i,k) \triangleq$ set of aircraft scheduled to land at i in period k

$E[W_k^i] \triangleq$ expected waiting time for an aircraft arriving to airport i at end of period k

$\lambda_k^i \triangleq$ number of scheduled arrivals at airport i during period k

The arrival times t_m^v are real numbers which represent times within the integer time periods. Time $t=0$ is the start of the operating day. Let $\kappa(t)$ be the function which takes real time values into their corresponding periods:

$$\kappa(t) = \lceil t/(\Delta t) \rceil .$$

The scheduled arrival rates $\{\lambda_k^i\}$ are determined from the sets of aircraft $\mathcal{A}(i,k)$ which are in turn determined by the itineraries $\mathcal{I}(v)$:

$$\lambda_k^i = |\mathcal{A}(i,k)| \tag{5.1}$$

$$\mathcal{A}(i,k) = \{v : (i,t,s) \in \mathcal{I}(v) \text{ for some } s \text{ and } \kappa(t) = k\} \tag{5.2}$$

Consider an aircraft which arrives at airport i at some time t during period k. A reasonable estimate of this aircraft's waiting time to land is the convex combination of expected waiting times at the end of periods $k-1$ and k,

$$\alpha E\left[W_{k-1}^i\right] + (1-\alpha)E\left[W_k^i\right], \tag{5.3}$$

with the weight α determined by whether t lies closer to the end of period k or $k-1$:

$$\alpha = \kappa(t) - t/(\Delta t).$$

Not all of this delay is necessarily propagated to later points in the system, however, because of slack, and cumulative delay d^v is adjusted to reflect this fact. To illustrate, let the above aircraft's next scheduled stop (stop $m+1$) be i' at time t', and suppose that from the current stop until the next stop there is an available slack of s'. Prior

to the mth stop, the aircraft's cumulative delay was d^v; thus its new scheduled arrival time is given by

$$t' + \left(d^v + \alpha E\left[W^i_{k-1}\right] + (1-\alpha)E\left[W^i_k\right] - s'\right)^+$$

In words, the aircraft's delay into its next stop is the maximum of zero and the value

$$X = \text{current delay} + \text{new congestion delay} - \text{schedule slack} .$$

Algorithm 1, based on this simple idea, is given in Figure 5.1.

The algorithm consists of two main parts: computation of expected waiting times and updating of schedules. To accomplish the former, we must aggregate aircraft and compute the level of demand at each airport, while in the updating procedure we must disaggregate again to the level of individual aircraft. Because of this repeated aggregation and disaggregation, the choice of data structures is important. For the implementation discussed here, the central data structure is the one illustrated in Figure 5.2. The arrival sets $\mathcal{A}(i,k)$ are singly linked lists of aircraft indexed by their currently scheduled destination i and arrival period k. Each aircraft record contains a pointer to its schedule, another linked list, and a pointer to the current destination in that schedule. In a given period, the number of aircraft records hung from a particular location in the data structure constitutes the demand rate. This counting is the aggregation procedure. Once the resulting queueing delay is calculated for this period and location, each affected aircraft record is re-hung from a new part of the arrival matrix based upon its slack and schedule. This update is the disaggregation procedure.

This choice of data structure means that the inner updating loop (the disaggregation procedure) requires only $O(V)$ time. The bottleneck of the algorithm consists of repeated calls to the subroutine for computing expected waiting times. Hence the following theorem on computational complexity.

Theorem 5.1 *The complexity of the expectation-based decomposition algorithm is $O(KNU)$, where U is the complexity of the single hub recursive algorithm for waiting*

First Decomposition Algorithm for Air Network Congestion

Initialize:
For $k = 1$ to K
  For $i = 1$ to N
    $\mathcal{A}(i,k) = \phi$
**** first itinerary stops are deterministic since not affected by earlier delays *****
For $i = 1$ to N
  For $v = 1$ to V
    $\mathcal{A}(i, \kappa(t_1^v)) = \mathcal{A}(i, \kappa(t_1^v)) \cup v$
Set $d^v = 0 \ \forall \ v$.

Main loop:
For $k = 1$ to K
  For $i = 1$ to N
    Set $\lambda_k^i = |\mathcal{A}(i,k)|$.
    Using the recursive method at each airport, calculate $E\left[W_k^1\right], \ldots, E\left[W_k^N\right]$.
    For $v \in \mathcal{A}(i,k)$:
    ***** find the part of the itinerary corresponding to this stop *****
      Find $m : (i_m^v, t_m^v, s_m^v) \in \mathcal{I}(v)$ and $\kappa(t_m^v + d^v) = k$
      Set $i = i_m$, $t = t_m + d^v$, $s = s_m$, $i' = i_{m+1}$, $t' = t_{m+1}$, $s' = s_{m+1}$.
      Set $\alpha = \kappa(t) - t/(\Delta t)$.
    ***** calculate propagated delay *****
      Set $d_{m+1}^v = \left[d^v + \alpha E\left[W_{\kappa(t)-1}^i\right] + (1-\alpha)E\left[W_{\kappa(t)}^i\right] - s'\right]^+$.
    ***** determine next arrival period and update data structure *****
      Set $\mathcal{A}(i', \kappa(t' + d^v)) = \mathcal{A}(i', \kappa(t' + d^v)) \cup v$.
END.

Figure 5.1: Decomposition algorithm for network based on deterministic updating scheme

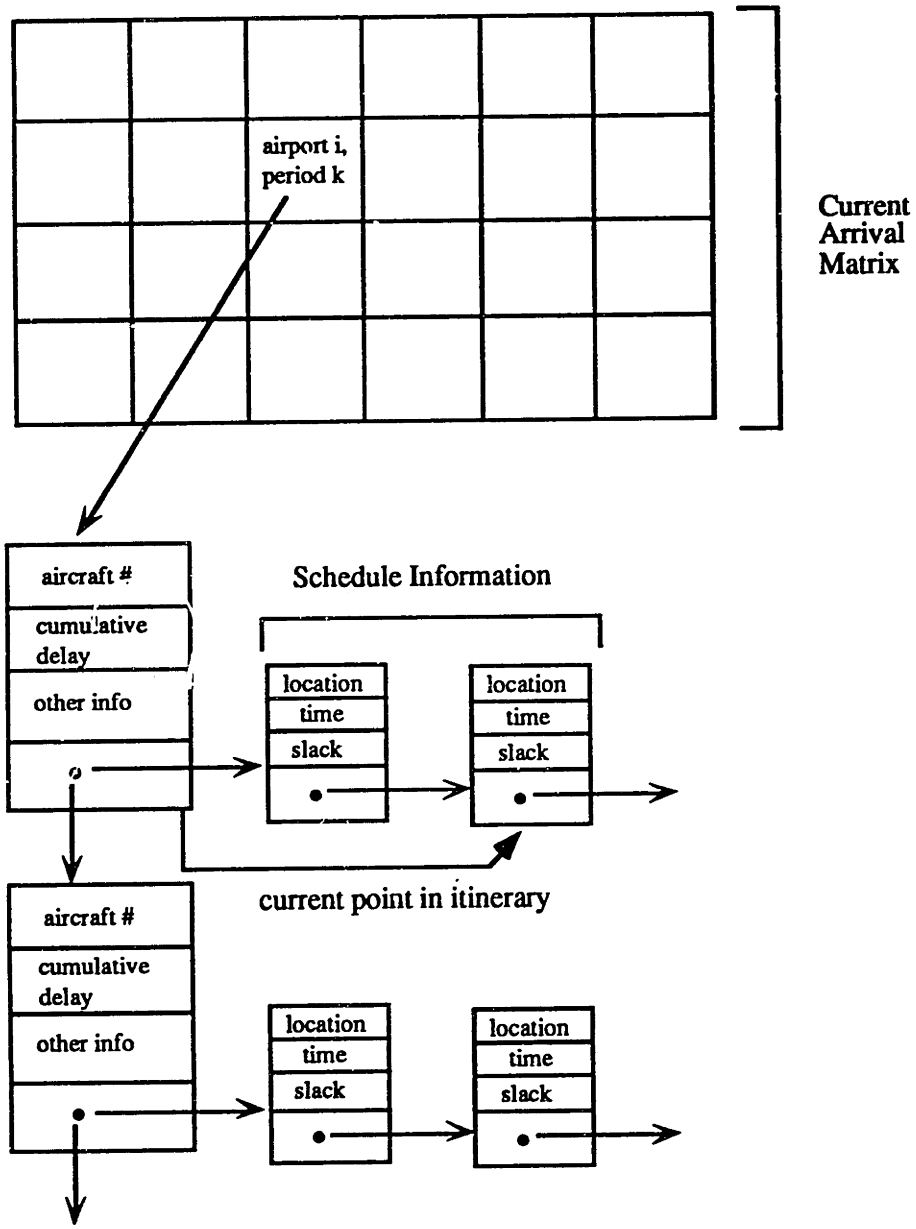


Figure 5.2: Data structure used in network congestion algorithms

time moments with deterministic input. If the Markov capacity model is specified with S capacity states, overall complexity is $O(NS^2K^3Q_{\max})$. □

The presence of the additional factor K in the complexity is due to the fact that at each time period k, the algorithm re-calculates all expected waiting times through period k. This duplication of effort could theoretically be reduced if it were possible to store the end conditions of iteration k to be used as initial conditions for iteration $k+1$. However, even for the simpler Markov capacity model, this would mean storing the joint probabilities for queue length and capacity. Since computing these probabilities requires $O(Q_{\max})$ times as much effort as for the expectation alone, there is no benefit to doing so unless they are desired for some other reason.

A more practical improvement is to have the recursion re-start only every m periods, where m is the minimum number of periods any aircraft has between scheduled stops. Under this scheme, the algorithm is run for the first m periods, arrivals are updated, then the algorithm is run for the first $2m$ periods, and so on. Whereas in the original implementation, the number of iterations performed within the recursive algorithm is

$$1 + 2 + \ldots + K = K(K+1)/2,$$

under this new scheme it is

$$m + 2m + 3m + Gm + K' = G(G+1)m/2 + K'$$

where $G = \lfloor K/m \rfloor$ and

$$K' = \begin{cases} K & \text{if } Gm < K \\ 0 & \text{otherwise} \end{cases}$$

This modification alone leads to substantial savings. The number of iterations is reduced by a factor

$$\begin{aligned} \frac{K(K+1)/2}{m\lfloor K/m \rfloor (\lfloor K/m \rfloor + 1)} &\geq \frac{K(K+1)}{K(K/m+1)} \\ &= \frac{K+1}{K/m+1.} \end{aligned}$$

In the case $K = 80$, for example, a value of $m = 10$ implies that the number of iterations is reduced from 3240 to 360, one-ninth of the former number. We note that because of the higher computational requirements of the network problem, the speed advantage of the Markov model over the semi-Markov model is substantial. All of the subsequent runs employ the Markov formulation.

5.1.2 An Algorithm with Probabilistic Updating

The updating scheme of the previous section takes deterministic arrival streams and uses expected waiting time information to convert them into new deterministic arrival streams. A more sophisticated method is suggested by the stochastic input refinement discussed in Section 3.2.3. The aim is to allow the variance in the waiting times, as well the as the means, to specify information about future arrival rates. These arrival rates are specified probabilistically rather than deterministically.

Consider a particular airport i at period k, and let the expectation and variance of the waiting time at that point be denoted simply as μ and σ^2. Suppose that it is possible from these parameters to estimate an approximate density $f_k^i(w)$ for the waiting time W_k^i. From this density and knowledge of the schedule slacks, one can then characterize (probabilistically) the next arrival period of each aircraft $v \in \mathcal{A}(i,k)$. More specifically, one can specify numbers $p_v(0), \ldots, p_v(C)$ and $k_v(0), \ldots, k_v(C)$ with the following interpretation: the next period in which aircraft v will have a landing is $k_v(0)$ with probability $p_v(0)$, $k_v(1)$ with probability $p_v(1)$ and so on up to some practical bound C. Figure 5.3 illustrates this phenomenon of traffic "splitting."

In order to complete the updating scheme, the algorithm must translate the probabilistic information on individual aircraft into information on future arrival rates. Defining the stochastic arrival quantities

$$\Lambda(i,k) \triangleq \text{number of arrivals at airport } i \text{ in period } k,$$

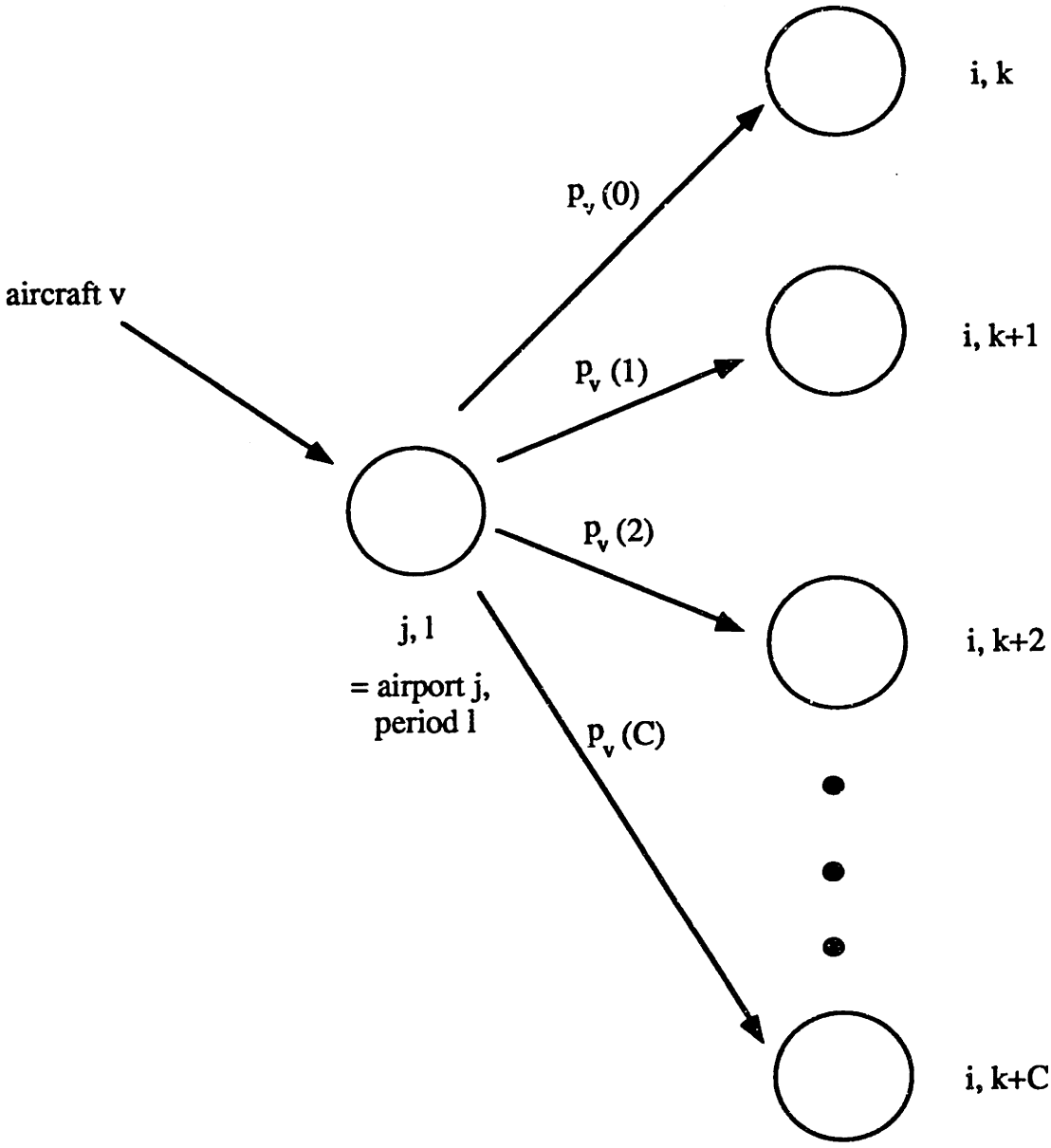


Figure 5.3: The traffic splitting phenomenon: alternative future aircraft paths depend upon delay encountered. The numbers $\{p_v\}$ indicate probabilities.

the algorithm estimates numbers $\gamma_k^i(1), \ldots, \gamma_k^i(L)$ and $\lambda_k^i(1), \ldots, \lambda_k^i(L)$ which describe the arrival process probabilistically:

$$\begin{aligned} \Pr\left\{\Lambda(i,k) = \lambda_k^i(1)\right\} &= \gamma_k^i(1) \\ \Pr\left\{\Lambda(i,k) = \lambda_k^i(2)\right\} &= \gamma_k^i(2) \\ &\vdots \\ \Pr\left\{\Lambda(i,k) = \lambda_k^i(L)\right\} &= \gamma_k^i(L). \end{aligned} \tag{5.4}$$

This simplified description of variability in the arrival rates is easily incorporated into the recursion for expected queue lengths, as was shown in Section 3.2.3. The recursion then produces future waiting time estimates, leading to new densities, new arrival probabilities, and so on.

An extremely important point is suggested by Figure 5.3. Because of uncertainty in delays, an aircraft landing at a particular place and time takes one of many future paths. Ideally, we would like to keep track of all such future paths and thus be able to assign probabilities to all realizations of the sets $\mathcal{A}(i,k)$. Unfortunately, the computational complexity inherent in this task is overwhelming because of the large number of such paths — $O(C^{r(v)})$ for each aircraft v, where $r(v)$ is the number of points in v's itinerary. Thus while we can reflect the splitting phenomenon in assigning probabilities to the different values $\lambda_k^i(\cdot)$, we must limit the realizations of the sets $\mathcal{A}(i,k)$. To accomplish this, we repeat the method of Algorithm 1, updating each aircraft's *cumulative delay* by a convex combination of $E[W_k]$ and $E[W_{k-1}]$. Thus, unlike Algorithm 1, Algorithm 2 allows a partial modeling of the splitting phenomenon (through the λ_k^i's), but it also introduces a potential inconsistency between schedule adjustment (traffic splitting disallowed) and demand rate adjustment (traffic splitting allowed). This inconsistency could adversely affect the results.

In total, the second decomposition algorithm requires four separate procedures:

1. Estimation of the densities $f_k^i\left(w; \mu(i,k), \sigma^2(i,k)\right)$ for the waiting times at each

station and period, given the estimates of mean and variance computed in the recursion.

2. Translation of these density functions into probabilistic descriptions of future arrival periods for each aircraft, as given in the parameters $p_v(0), \ldots, p_v(C)$ and $k_v(0), \ldots, k_v(C)$.

3. Translation of the individual aircraft parameters $p_v(0), \ldots, p_v(C)$ and $k_v(0), \ldots, k_v(C)$ into simple discrete distributions for the random variables $\Lambda(i, k)$.

4. Updating of aircraft itineraries and airport arrival lists.

The fourth of these procedures was described in the previous subsection. The first three are described in further detail in what follows, and a summary of the algorithm is given in Figure 5.6.

Obtaining waiting time densities

Estimation of the densities $f(w)$ cannot be done on the basis of the recursive algorithm alone, since this procedure gives only the first two moments of the distribution. Knowledge of the third moment would give enough information to determine a unique 2-point discrete distribution by solving the non-linear system

$$\begin{aligned} p_1 w_1 + p_2 w_2 &= E[W] \\ p_1 w_1^2 + p_2 w_2^2 &= E\left[W^2\right] \\ p_1 w_1^3 + p_2 w_2^3 &= E\left[W^3\right] \\ p_1 + p_2 &= 1 \\ p_1, p_2, w_1, w_2 &\geq 0. \end{aligned} \tag{5.5}$$

for the values p_1, p_2, w_1, and w_2. However, this system is not guaranteed to have any solution because of the positivity requirement.

An alternative method is suggested by the discussion of Section 3.3, in which we proposed a 2-parameter distribution for waiting times based on simulation results. The distribution is reproduced here as [c.f. (3.23)]

$$\begin{aligned} \Pr\left\{W_k^i = w_{\min}(i,k)\right\} &= \delta \\ \Pr\left\{W_k^i \le w \mid w > w_{\min}(i,k)\right\} &= 1 - e^{-\nu(w - w_{\min})}. \end{aligned} \tag{5.6}$$

Recall from the earlier discussion that the parameter $w_{\min}$ is determined directly within the recursion, while the parameters δ and ν must be determined by solving a pair of equations [equations (3.24)] using the first two waiting time moments. In terms of the mean $\overline{w}$ and variance σ^2 we obtain the solution (omitting subscripts)

$$\delta = \frac{\sigma^2 - (\overline{w} - w_{\min})^2}{\sigma^2 + (\overline{w} - w_{\min})^2} \tag{5.7}$$

$$\nu = \frac{2(\overline{w} - w_{\min})}{\sigma^2 + (\overline{w} - w_{\min})^2} \tag{5.8}$$

Note that δ is always less than 1 and will be nonnegative provided that

$$\frac{\sigma^2}{(\overline{w} - w_{\min})^2} \ge 1.$$

In the typical case where $w_{\min}$ is zero, this is equivalent to the condition that the coefficient of variation exceed 1. Only in rare instances of the tests presented shortly was this condition found not to hold. In those cases, the parameter δ was set to 0 and the entire distribution assumed to be exponential.

From densities to schedules

Given estimated densities for W_k^i for all points i in the network, the next step in the procedure is to infer probabilities for the immediate future paths of all aircraft $v \in \mathcal{A}(i,k)$. For any such aircraft, let (i', t', s') be the scheduled next stop (stop $m+1$) on its itinerary. The earliest period in which this aircraft's next landing may actually take place is

$$k_v(0) = \kappa\left(t' + [d^v + w_{\min} - s']^+\right).$$

This is the earliest period at which this aircraft could next land, reflecting the minimum waiting time achievable at this stop (usually 0). Accordingly, the greatest amount of delay this aircraft can endure at i and have this next arrival period remain unaltered is

$$\begin{aligned} w(0) &= \max\{w' : \kappa(t' + [d^v + w' - s']^+) = k_v(0)\} \\ &= \{w' : t' + d^v + w' - s' = k_v(0)\Delta t\} \\ &= k_v(0)\Delta t - t' - d^v + s'. \end{aligned}$$

where d^v is its cumulative delay prior to the mth stop. The probability that the aircraft's next scheduled period is $k_v(0)$ is

$$p_v(0) = \int_{w_{\min}}^{w(0)} f(w; \mu, \sigma^2) dw. \tag{5.9}$$

If $w_{\min} = 0$, which is usually the case, $k_v(0)$ corresponds to the outcome that zero additional periods of delay are added to aircraft v at this stop. When the waiting time density is approximated by the distribution (5.6) with $w_{\min} = 0$, (5.9) becomes

$$p_v(0) = \delta + (1 - \delta)\left[1 - \exp(-\lambda w(0))\right].$$

Letting $w(1) = w(0) + \Delta t$, the probability of the next scheduled period being $k_v(1) \equiv k_v(0) + 1$ is

$$p_v(1) = \int_{w(0)}^{w(1)} f(w; \mu, \sigma^2) dw, \tag{5.10}$$

and in general the probability of c additional periods of delay is

$$p_v(c) = \int_{w(c-1)}^{w(c)} f(w; \mu, \sigma^2) dw, \tag{5.11}$$

where $w(c) = w(0) + c\Delta t$. These have the appropriate specific forms when the distribution (5.6) is substituted.

For practical reasons, it is necessary to choose some upper bound C on the number of periods delay to allow. Hence

$$p_v(C) = \int_{w(C-1)}^{\infty} f(w; \mu, \sigma^2) dw,$$

Together with the numbers $\{k_v(c)\}$, the probabilities $\{p_v(c)\}$ then constitute a probabilistic description of the next period in which aircraft v will demand to land.

Characterizing arrivals

In order to translate the numbers $\{p_v(c)\}$ into a probabilistic description of the future demand rates $\Lambda(i,k)$, define the random variable

$$X_{jlik}(v) \triangleq \begin{cases} 1 & \text{if } v \in \mathcal{A}(j,l) \text{ is delayed such that its} \\ & \text{next stop will be } i \text{ at period } k \\ 0 & \text{otherwise} \end{cases}$$

This random variable denotes the "contribution" of an arrival at one place and time to the arrival rate at a future place and time. Note that if the next stop of $v \in \mathcal{A}(j,l)$ is i, then

$$\Pr\{X_{jlik}(v) = 1\} = p_v(k-l).$$

In words, for aircraft $v \in \mathcal{A}(j,l)$, the probability that it will contribute to the landing demand at airport i during period k (assuming that j is its next scheduled stop) is $p_v(k-l)$.

The random variables $X_{jlik}(v)$ provide the necessary connection between aircraft and arrival rates. A moment's thought shows that

$$\Lambda(i,k) = \sum_{j=1}^{N} \sum_{l<k} \sum_{v=1}^{V} X_{jlik}(v). \tag{5.12}$$

In words, this says that the arrival rate at (i,k) is the sum of all contributions from previous points in the itineraries (see Figure 5.4). Note the form of (5.12). The random variables $\{\Lambda\}$ are sums of Bernoulli random variables. Making the definition

$$NL(v,k) \triangleq \text{next destination of aircraft } v \text{ after period } k$$

the expectation is easily obtained as

$$\begin{aligned} E[\Lambda(i,k)] &= \sum_{j=1}^{N} \sum_{l<k} \sum_{v=1}^{V} E[X_{jlik}(v)]. \\ &= \sum_{j=1}^{N} \sum_{l<k} \sum_{v:NL(v,l)=i} p_v(k-l) \end{aligned} \tag{5.13}$$

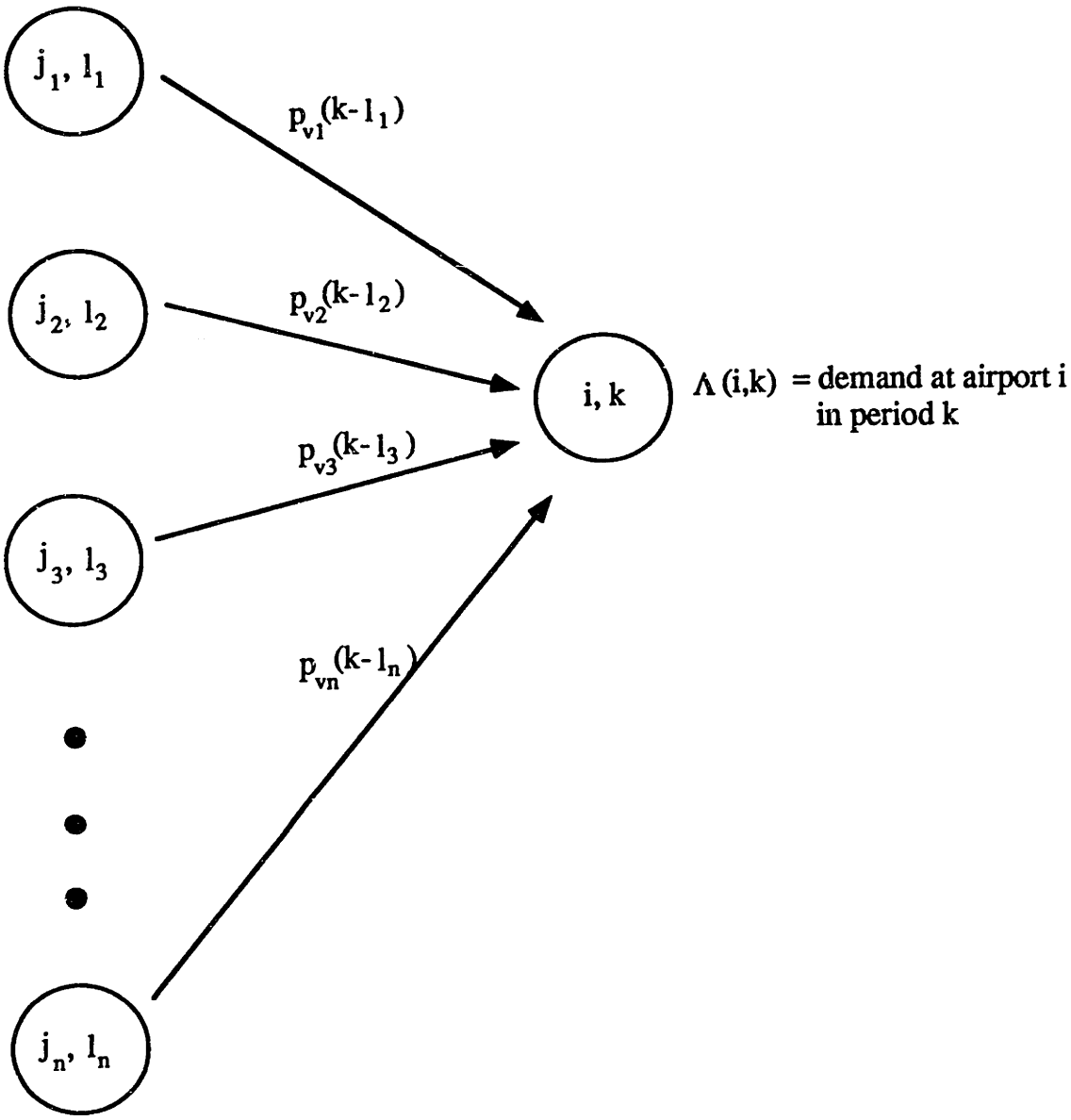


Figure 5.4: Updating downstream arrivals in Algorithm 2: early arrivals and delays contribute to demands later in the day.

Obtaining variances is not straightforward because the terms of the sum are not independent. Aircraft delayed at earlier points in the day may share the same source for those delays, so that their contributions to future demands may be correlated. On the other hand, diversity in scheduling and slack weaken this dependence. For the sake of tractability, we make the approximation that the contributions are approximately independent and write

$$Var[\Lambda(i,k)] \approx \sum_{j=1}^{N} \sum_{l<k} \sum_{v:NL(v,l)=i} p_v(k-l)(1-p_v(k-l)). \tag{5.14}$$

Simulation results indicate that this approximation is fairly good.

The specification of approximate distributions for the $\{\Lambda(i,k)\}$ is the final step in translating aircraft delays into arrival rate information. If we could compute the third moment, we could determine a 2-point distribution by solving a non-linear system such as (5.5). However, there is no straightforward way to obtain a third moment (which reflects skewness) other than simulation. An alternative is to assume a normal form for Λ and discretize into a suitable number of points. Such a normality assumption has some basis in the central limit theorem, but convergence may not be good because of non-independence between terms of the sum. Simulation results indicate that for early periods of the day where there are fewer terms in the sum, strange skewness patterns are possible (see Figure 5.5). These patterns disappear later in the day. While this phenomenon is some cause for concern, the test runs indicate a degree of insensitivity to the demand rate distribution. We retain the normality assumption while acknowledging its imperfections.

5.1.3 Complexity, Model Power, and Perspective

Although Algorithm 2 involves considerably more modeling work than Algorithm 1, its computational complexity is not significantly higher. Within the principal loop, the bottleneck operation remains that of calculating the waiting time moments in the recursion. Because the arrival stream is specified probabilistically rather than

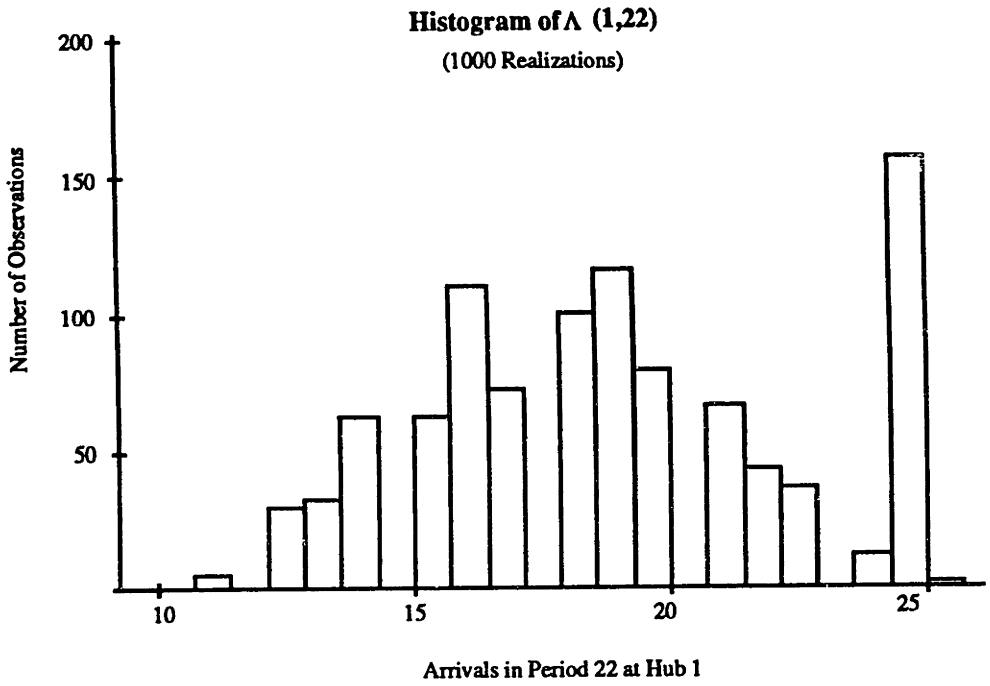


Figure 5.5: Histogram of $\Lambda(1,22)$ obtained from simulation. Unusual skewness patterns such as this one may occur when the contributing prior arrivals are still largely deterministic.

deterministically, there is an additional factor R equal to the number of values specified for each arrival rate distribution. Hence we have the following theorem for the computational complexity of Algorithm 2:

Theorem 5.2 *The complexity of the Algorithm 2 is $O(RKNU)$, where U is the complexity of the single hub recursive algorithm for waiting time moments with deterministic input. If the Markov capacity model is specified with S capacity states, overall complexity is $O(RNS^2K^3Q_{\max})$.* □

Both Algorithms 1 and 2 are suitable for any kind of network. However, their running times are not trivial: for a simple 2-airport network with $K = 80$ periods at each airport, Algorithm 1 takes about one hour on a DEC-3100 workstation while Algorithm 2 takes about three hours ($R=3$). Note that these running times do *not* reflect the reduction in calls to the recursion discussed earlier. Even with

Second Decomposition Algorithm for Air Network Congestion

Initialize:

For $k = 1$ to K

For $i = 1$ to N

$\mathcal{A}(i,k) = \phi$

$E[\Lambda(i,k)] = 0$

$\sigma^2[\Lambda(i,k)] = 0$

**** First demand period for each aircraft is fixed *****

For $v = 1$ to V

$\mathcal{A}(i,\kappa(t_1^v)) = \mathcal{A}(i,\kappa(t_1^v)) \cup v$

For each $(i,t,s) \in \mathcal{I}(v)$, $E[\lambda^i_{\kappa(t)}] = E[\lambda^i_{\kappa(t)}] + 1$

Set $d^v = 0 \ \forall \ \ v$.

Main loop:

For $k = 1$ to K

For $i = 1$ to N

From $E[\Lambda(i,k)]$ and $\sigma^2[\Lambda(i,k)]$ determine the quantities

$\lambda^i_k(1), \ldots, \lambda^i_k(L)$ and $\gamma^i_k(1), \ldots, \gamma^i_k(L)$

Using the recursive algorithm with probabilistic input λ, μ

calculate $E\left[W_k^1\right], \ldots, E\left[W_k^N\right]$ and $\sigma^2\left(W_k^1\right), \ldots, \sigma^2\left(W_k^N\right)$

**** Update itineraries — same way as first algorithm *****

For $v \in \mathcal{A}(i,k)$:

Find $m : (i_m^v, t_m^v, s_m^v) \in \mathcal{I}(v)$ and $\kappa(t_m^v + d^v) = k$

Set $i = i_m$, $t = t_m + d^v$, $s = s_m$, $i' = i_{m+1}$, $t' = t_{m+1}$, $s' = s_{m+1}$.

Set $\alpha = \kappa(t) - t/(\Delta t)$.

Set $d^v_{m+1} = \left[d^v + \alpha E\left[W^i_{\kappa(t)-1}\right] + (1-\alpha)E\left[W^i_{\kappa(t)}\right] - s'\right]^+$.

Set $\mathcal{A}(i', \kappa(t' + d^v)) = \mathcal{A}(i', \kappa(t' + d^v)) \cup v$.

**** Update future arrival rates *****

From α, $E[W^i_k]$, and $\sigma^2(W^i_k)$, determine the densities $\{f^i_k(w)\}$.

From the densities $f^i_k(w)$, determine the quantities

$p_v(0), \ldots, p_v(C)$ and $k_v(0), \ldots, k_v(C)$ $\forall v \in \mathcal{A}(i,k)$.

For $c = 0$ to C:

$E\left[\Lambda^i_{k(v,c)}\right] = E\left[\Lambda^i_{k(v,c)}\right] + p_v(c)$

$\sigma^2\left(\Lambda^i_{k(v,c)}\right) = \sigma^2\left(\Lambda^i_{k(v,c)}\right) + p_v(c)(1 - p_v(c))$

END

Figure 5.6: Decomposition algorithm for network based on stochastic update scheme

this factor-ten improvment, however, modeling a full-size network of a large airline (400+ nodes) is a daunting problem. On the other hand, the problem is well-suited to parallel computation, with different processors handling the individual nodes and a central processor controlling the bookkeeping of aggregation and disaggregation. Without such parallel capability, one must ask the question of whether it is necessary to model all airports in the network or whether further simplification is possible.

Consider a single carrier trying to understand congestion in its own hub-and-spoke network. From this carrier's perspective, delays at its *hubs* have far greater implications for disruption of its schedule than delays at its *spokes*. This observation suggests a simplification: reduce the hub-and-spoke network to a network of hubs. That is, keep track only of aircraft belonging to the hub carrier, treat other arrivals as fixed, and treat all congestion delays other than those emanating from the hubs as *negligible*. In the resulting collapsed network, we incorporate spoke information in setting aircraft itineraries. As before, these consist of ordered triples $\{(i_m, t_m, s_m)\}$, but now each i_m refers to a hub airport and each s_m reflects the total slack available to an aircraft between successive visits to hubs, including the slack available at an intervening spoke. External aircraft add to demand and congestion in the system, but their arrival schedules are considered fixed. All internal flights in the collapsed network appear to take place between hubs, but flight times vary widely to reflect the fact that in reality, the aircraft have intermediate spoke stops.

By ignoring congestion at the spokes of the system and concentrating only on the hubs, we reduce the size of a large airline's network from 400+ nodes to perhaps 5 or 6. These changes reduce the model's realism, but the reduced model should capture essential behavior. Since one of the main goals is to improve our understanding of interactions between *hubs* (e.g. the issue of isolating Chicago), this simplification seems to be further justified. The testing and analysis presented in the remainder of this chapter is conducted on a simple 2-hub network which embodies these ideas.

5.2 Testing the Decomposition Models

It is desirable to test the validity of the results obtained from the network decomposition models, much as the Dallas case study examined the usefulness of the single-hub model. However, rather than consider actual data as we did in Chapter 4, we present here results of a small "case study" conducted on a simple hypothetical 2-hub network, using demand and capacity data which resemble those found in practice. Section 5.2.1 discusses the set-up of the test procedure, including the issues to be addressed, while Section 5.2.2 presents a discussion of our findings.

5.2.1 The Testing Procedure

The following set of questions will guide our test procedure:

- How well do the network approximation procedures work?
- What is the effect of congestion at one hub on demand and congestion at the other?
- How is this effect altered by the amount of slack in aircraft schedules?
- What is the effect of isolating a congested hub by not allowing its flights to connect with the other hub?

Figure 5.7 illustrates the form of our test network. Along the lines of the above remarks, we simplify to create a network of two hubs. The test runs have two sources of landing demand at each hub: external demands $\{\nu_k^i\}$, which are specified as parameters, and internal demands $\{\lambda_k^i\}$ which are endogenously determined according to schedules and delays.

Table 5.1 summarizes the main characteristics of the various test cases. The cases differ with respect to the presence or absence of banks of flights, the degree of separation between banks, the amount of slack in aircraft schedules, the traffic intensity, the initial capacity conditions, and the percentage p of flights which visit

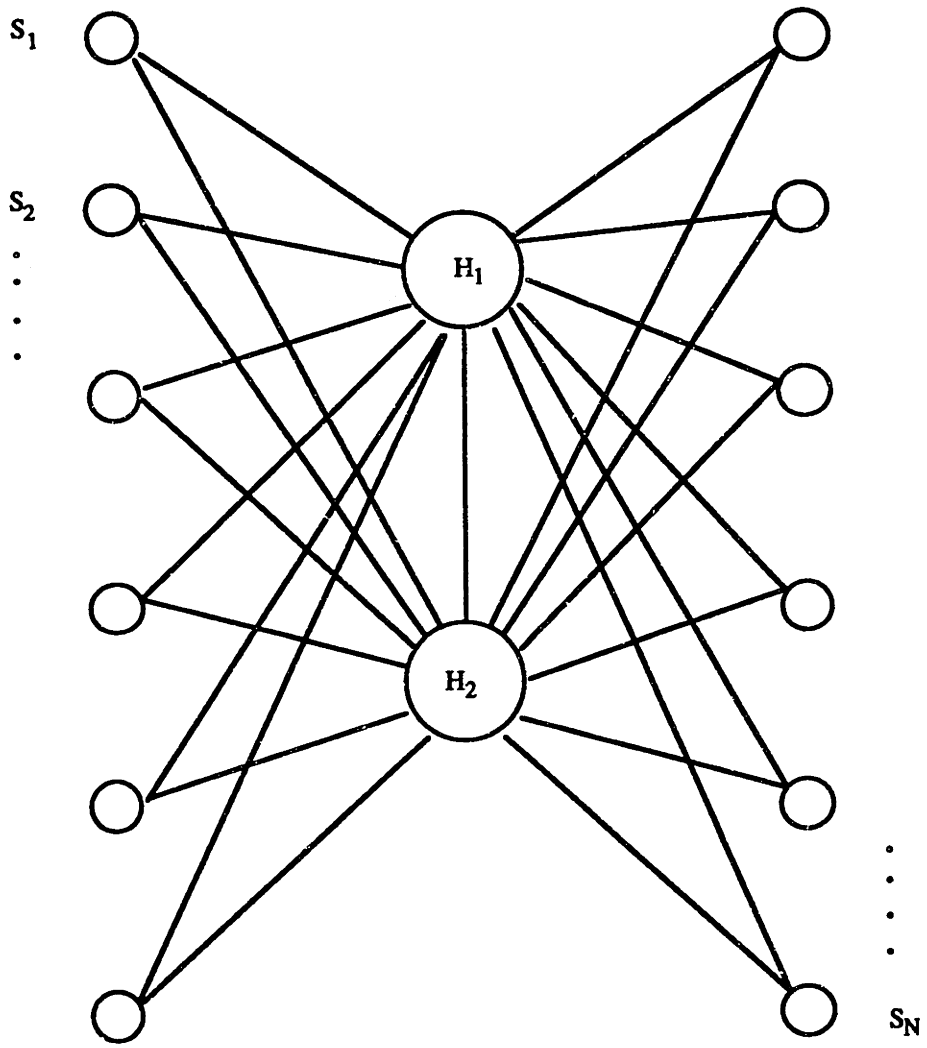


Figure 5.7: Schematic view of 2-hub network

case #	no. banks	bank space	p	ρ	slack	initial capacities
1	(DFW)	—	0	0.5	15-20 mins.	low/high
2a	12	15 mins.	0.5	0.9	5 mins.	steady state
2b	12	15 mins.	0.5	0.9	500 mins.	steady state
3	—	—	0.5	0.8	5 mins.	steady state
4a	10	30 mins.	0	0.7	5 mins.	low/high
4b	10	30 mins.	1	0.7	5 mins.	low/high
5a	10	30 mins.	0.5	0.7	5 mins.	steady state
5b	10	30 mins.	0.5	0.7	10 mins.	steady state
5c	10	30 mins.	0.5	0.7	15 mins.	steady state
5d	10	30 mins.	0.5	0.7	20 mins.	steady state

Table 5.1: Test run information. Note that traffic intensities ρ are based on that part of the schedule which does not include the runout period at the end of the day.

different hubs (rather than the same hub) on alternate visits. This latter statistic is a measure of how each hub is tied to the performance of the other. In the schedules, an aircraft with an arrival at a given hub H has its next hub arrival at the other location with probability p and at H with probability $1-p$. A value of $p = 1$ implies a fully connected network, while a 0 value means a totally disconnected network.

Validation: Cases 1,2, and 3

Cases 1,2, and 3 are concerned with validation of the network models. In each case we test the models against a simulation procedure like the one first introduced in Chapter 3. This simulation generates period-by-period capacities at each hub using Monte Carlo methods. It works in exactly the same fashion as the two approximation procedures, except that arrival rates are adjusted by simulated waiting times rather than by expected values or some approximate distribution of waiting time.

The demand and capacity data for case #1 closely resemble those at DFW. We have collapsed the capacity state space to the three states corresponding to 'A', 'D', and 'F' at Dallas. These have corresponding steady state probabilities 0.07, 0.10, and 0.83. The simulated demand, together with the actual DFW demand, is shown in Figure 5.8. Aircraft slack for this experiment takes values in the range of 15-20 minutes between stops at hubs, depending on the distance to the intervening spoke.

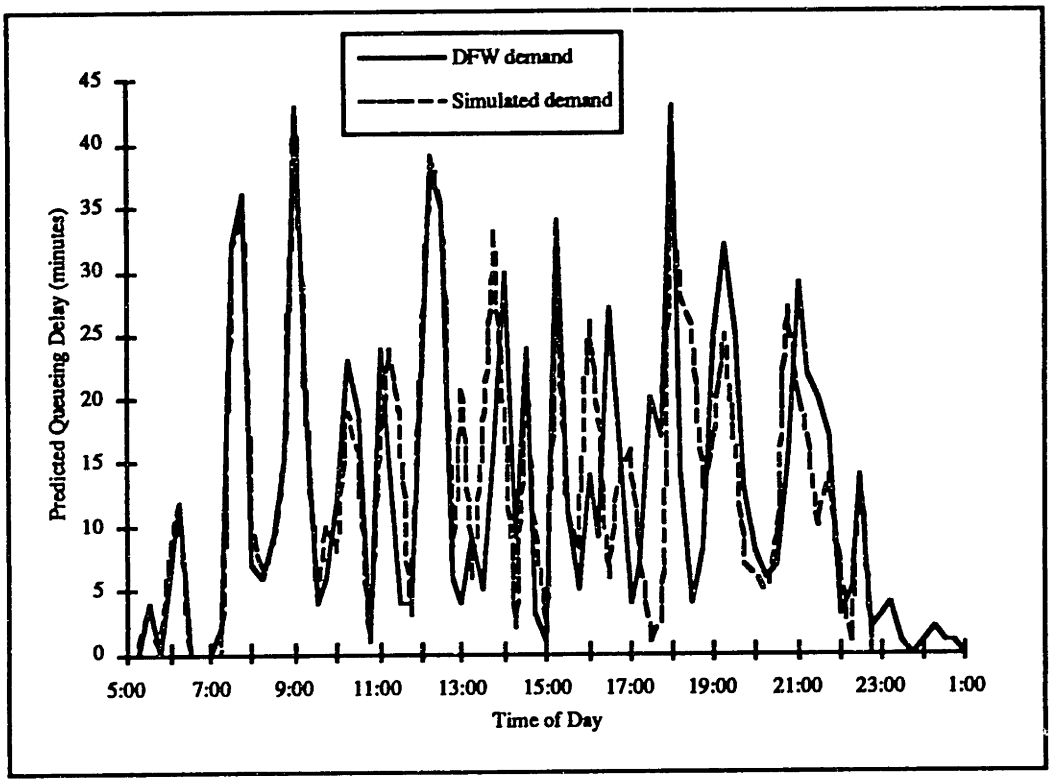


Figure 5.8: Case #1 has demand simulated to resemble DFW.

Case #2 has internal aircraft grouped into identically timed banks of 30-minutes duration at each hub with relatively short inter-bank periods of 15 minutes. Peak demands are higher than at DFW, capacities are slightly lower, and the underlying Markov chain has steady state probabilities $0.26, 0.21, 0.53$. The result is a considerably higher traffic intensity here than in case #1. Figure 5.9 illustrates the hypothetical arrival patterns. For this case we report results for runs with very low aircraft slack ($s = 5$ minutes) and artificially high slack ($s = 500$ minutes). The latter instance is investigated to ascertain the effect of hub interaction. This is an artificial experiment because the same schedule is retained as with $s = 5$, i.e. the 500-minute slack actually exceeds the time between visits. The effect is as if delays encountered are not propagated to later parts of the day. Comparison with the low-slack case thus illustrates the effect of delay propagation on the arrival schedules.

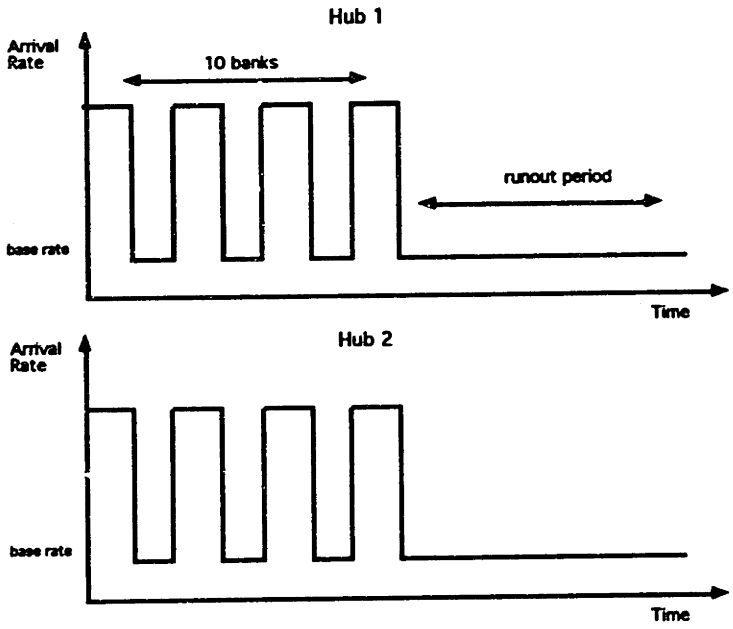


Figure 5.9: Shape of 2-hub hypothetical demand for cases 2, 4, and 5

Case #3 reports results for a continuous demand pattern at the two airports (no banks).

Effects of Slack and Connectivity: Cases 4 and 5

In cases 4 and 5 we are concerned with illustrating the effects of slack and network connectivity. Both experiments have a traffic pattern like that of #2 except that there is lower landing demand and greater separation (30 minutes) between banks. Case #4 illustrates the idea of hub isolation by considering an instance in which the two hubs have no aircraft in common ($p=0$, the disconnected case) and an instance in which they have all aircraft in common ($p=1$, the fully connected case). Case #5 considers four instances in which aircraft slack is varied. In each of these we are concerned with the resultant effects on *cumulative aircraft delays* rather than on expected waiting times at each airport over the course of the day.

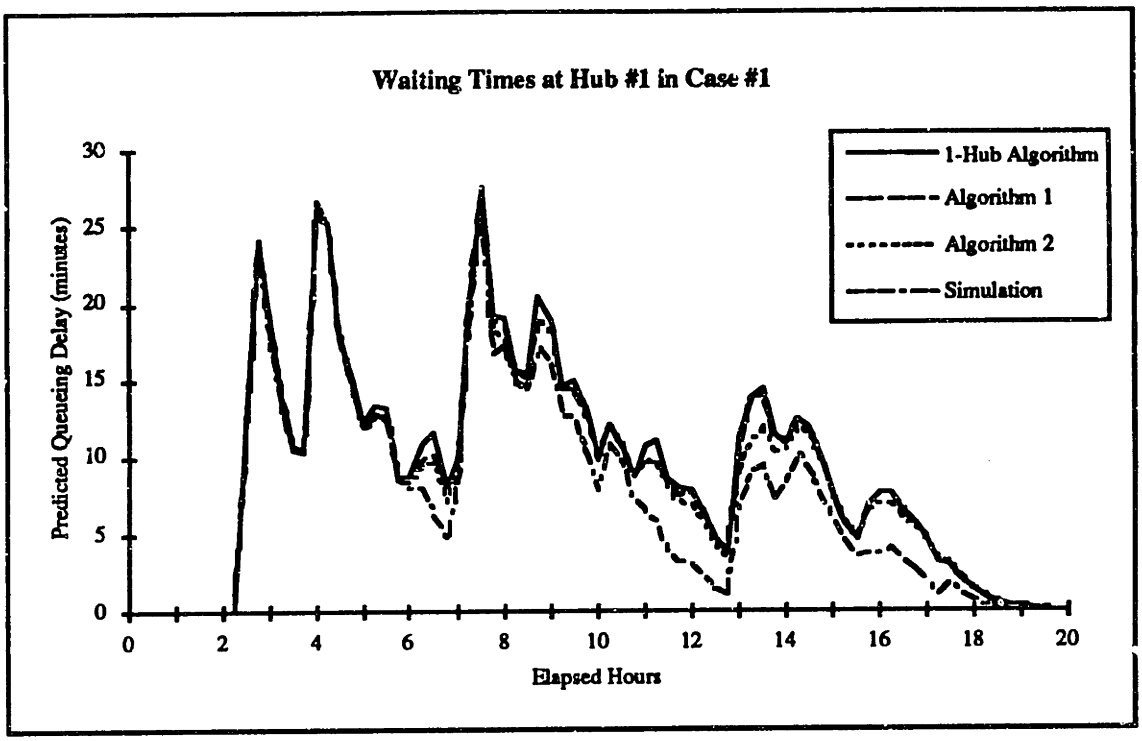


Figure 5.10: Comparison of expected waiting times predicted by one-hub algorithm, simulation, and the two decomposition algorithms for case #1

5.2.2 Results and Discussion

We remark at the outset that model parameters have a noticeable effect on the mechanics of the network. In the DFW case (#1), waiting times are of the same order as aircraft slack, and there is substantial separation between major traffic peaks. For these reasons, we expect delay propagation to be relatively low and have a less disruptive effect on the schedule. In case #2, on the other hand, major peaks are much closer together, traffic intensity is sharply higher, and delay propagation should be more important. Case #3 lies in between these first two cases.

Case #1: Comparison with Single-hub Algorithm

In case #1 we focus only on the first hub. Figure 5.10 plots expected waiting times at this hub as estimated by four different procedures: the single-hub algorithm of Chapter 3, decomposition algorithms 1 and 2, and simulation. There is fairly close

agreement between all four curves. The solid line indicates the predictions of the one-hub algorithm, in which all demands are treated as external and no account is made for propagation. The curve for Algorithm 1 (update by expected value) is quite close to this first curve, reflecting the fact that slack values (15-20 minutes) are approximately equal to expected waiting times and hence delay propagation is minimal. Algorithm 2 reflects the effects of delay propagation slightly more, though the differences are still small. Finally, with the simulation curve we see a further departure from the one-hub results.

The most striking features of the graph are the closeness of the four curves and the preservation of the peaked delay pattern, both of which indicate that the effects induced by delay propagation (the "network effects") are relatively minor. Because there is ample space between major banks and slack values are close to the mean waiting times, the general peaked pattern is preserved. However, as the next case illustrates, the situation changes when traffic becomes heavier and spacing between major banks is decreased.

Case #2: Effects of Heavier Traffic and Closer Spacing

A more difficult test for the network approximations is provided by case #2a (see Figure 5.11). Here expected waiting times (30-40 minutes) are quite high relative to aircraft slack (5 minutes), and there is only a 15-minute gap between successive banks. Thus, while the early part of the day shows a close fit between the simulation and the algorithms (the deterministic effect), there is a noticeable disparity in the middle part of the day, when alterations in the arrival stream become significant. Notice that relative to simulation, both algorithms tend to overestimate delay during the middle part of the day, with the error as high as 30% for certain periods.

For a given hub (1 or 2) and algorithm (1 or 2) we can define a standard error in the predictions relative to simulation. Let X_k denote the waiting time value predicted by algorithm for period k and Y_k denote the corresponding value for the

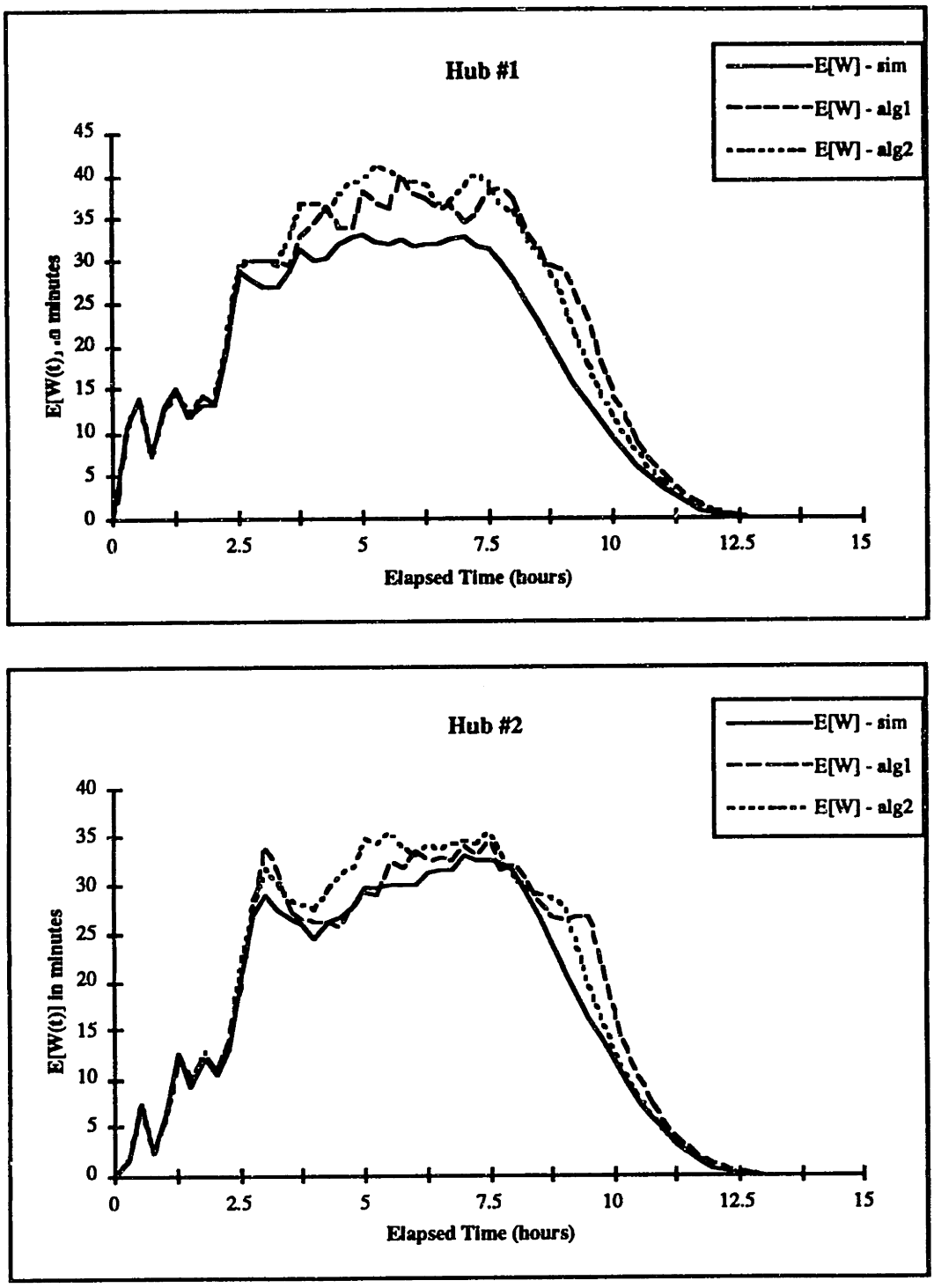


Figure 5.11: Comparison of expected waiting times predicted by simulation and the two decomposition methods for case #2a

simulation. Then the standard error s is given by

$$s = \sqrt{\frac{\sum_{k=1}^{K} (X_k - Y_k)^2}{K}}.$$

This provides a measure of how far apart the simulation and algorithm results are. For Algorithm 1, these values are 4.5 and 2.6 minutes at the two hubs, while the corresponding numbers for Algorithm 2 are 4.5 and 2.3. The numbers represent an average error of 10-20%, with worse fits in the middle part of the day.

The Demand-smoothing Phenomenon

Note how the waiting time profiles are much smoother in case #2a than in case #1. With only a 15-minute separation between banks, the relatively high waiting times combine with low aircraft slack to overwhelm the bank structure, as illustrated at the top of Figure 5.12. This figure plots the original demand profiles at Hub #1 with those which occur as a result of the smoothing action of delay (the peaked case is labelled $s = 500$, for reasons explained below). In the high-traffic, low-slack case, propagation effects smooth the demand pattern substantially, with large numbers of aircraft shifted to periods quite late in the day. The sharp peak structure of the original demand is considerably altered.

Demand smoothing lies at the heart of why Algorithms 1 and 2 consistently overestimate delays in the middle part of the day (a phenomenon noticed in all test runs). In the actual process, an aircraft scheduled at a given period may experience a delay ranging from zero up to a very long time, perhaps 3 hours or more. In cases of high waiting times, the aircraft's next arrival time will be considerably later than was scheduled, and its contribution to later demand is pushed back by a significant number of periods. Thus over a large number of simulations with heavy traffic, a noticeable fraction of arrivals are pushed back to the later part of the day, when there is no scheduled traffic. Because capacity is more than adequate then, the result of this traffic shift is to reduce waiting times.

Ideally, the computational algorithms should reflect this shifting and smoothing

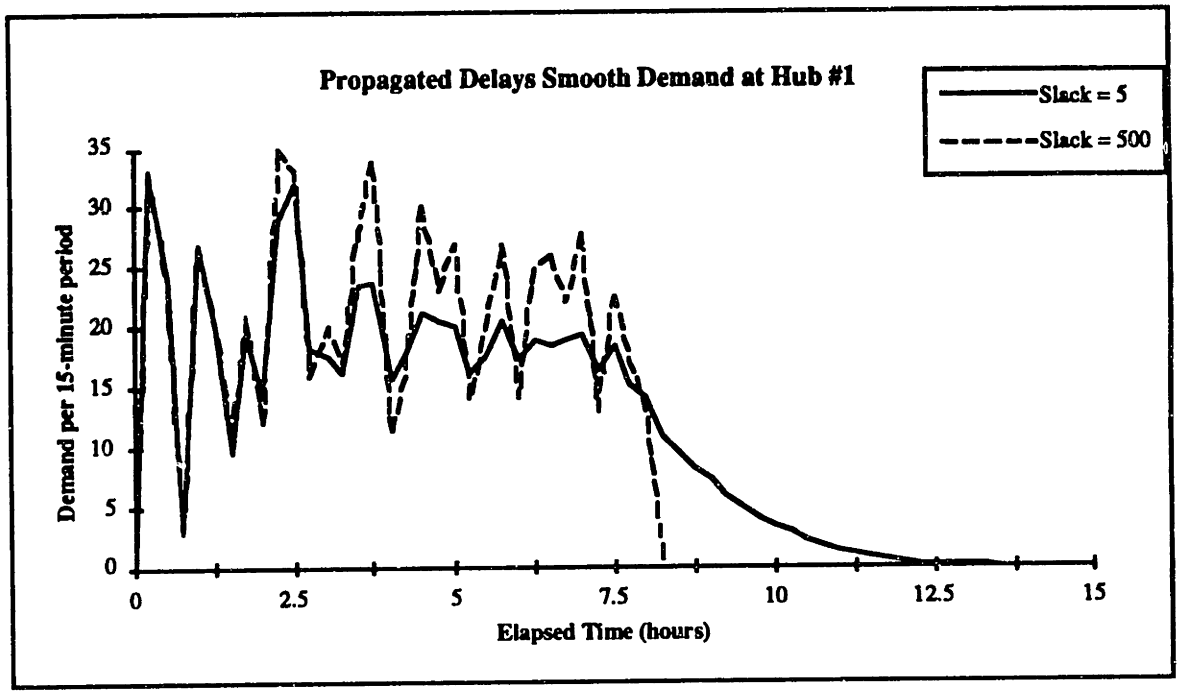
of demand. However, as was remarked earlier, to do a thorough job they would have to keep track of the thousands of potential paths which aircraft may follow as a result of delay, a seemingly impossible computational burden. To limit the state space to manageable size, both algorithms update aircraft schedules according to one number, expected waiting time. The result is that both algorithms tend not to shift aircraft to the very late part of the day but rather to concentrate demand more in the middle part, resulting in higher predicted waits.

Demand Smoothing and Delay

The phenomenon of demand shifting and smoothing explains other observations which seem counter-intuitive at first. An example of such a result is the fact that higher aircraft slack may actually *increase* expected queueing times at the hubs. Cases 2a and 2b illustrate. Both are cases of heavy traffic organized into narrowly separated banks. The difference is that in case #2b, each aircraft is given artificially high slack (500 minutes) while in case #2a aircraft have very low slack (5 minutes). In the 'b' case, the amount of slack actually exceeds the time between stops — it is an artificial case in which no matter what delay an aircraft encounters, it will reach its next destination on schedule. In a heavy traffic situation like case #2, this high slack acts to preserve the original concentrated demand pattern, whereas the low slack case allows the demand to become smoother over time as aircraft are pushed back to the end of the day (look again at the top of Figure 5.12). In the high slack case, the higher concentration of demand produces *higher* queueing delays, as we see in the bottom half of Figure 5.12. Here we also see a fairly close fit between the algorithms and the simulation, because the high slack removes the uncertainty inherent in the schedule disruption.

Case #3: Continuous Demand

Case #3 compares the performances of the algorithms with simulation when the demand is allowed to be continuous over the day (rather than organized into banks).



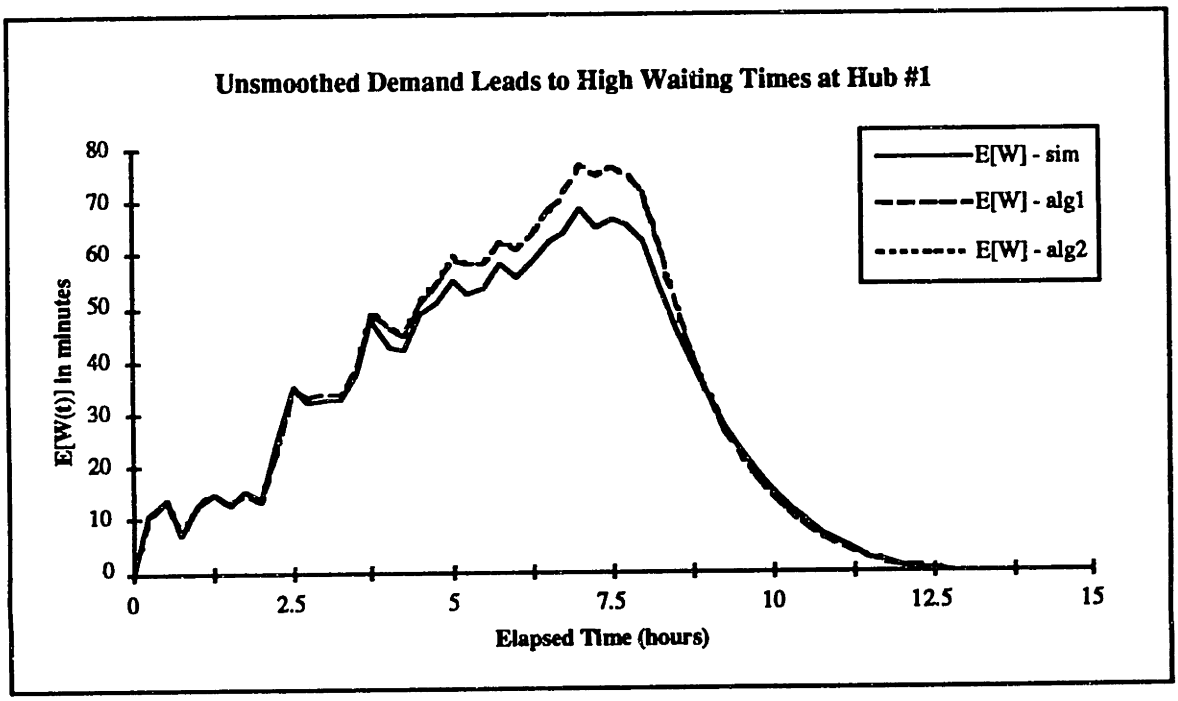


Figure 5.12: Top — demand rate at hub #1 with and without the smoothing effects produced by delays in the arrivals of incoming aircraft. Bottom — high waiting time at hub #1 produced by unsmoothed demand.

The results are illustrated in Figure 5.13. Again, the algorithms give estimates of waiting times greater than those predicted by simulation. The traffic intensity for this case is higher than case #1 but lower than case #2. The difference between the algorithms and simulation exceeds 20% for a large part of the day at hub 2, while the standard errors are approximately 15% of the delays: 2.2 minutes at hub 1 (for both algorithms) and 2.4 and 2.6 minutes (Algorithm 1 and Algorithm 2) at hub 2.

Summary of Cases 1,2, and 3

Cases 1,2, and 3 collectively suggest that the deterministic part of the schedule has a fairly large effect. This effect is most important when traffic is moderate and slack and bank separation are large, as in case #1. In cases of heavier traffic, lower slack, and less separation, the performance of the algorithms worsens as they tend not to capture the true spreading of demand. The results suggest that for a hub like DFW, network effects may be less important than for a case like Chicago's O'Hare, which has a more extended busy period.

Effects of Slack and Connectivity

The remainder of the cases (#4 and #5) illustrate the effect of network connectivity and aircraft slack on *cumulative aircraft delay*. This delay should be distinguished from waiting times at the hubs; the former is the sum of several instances of the latter, with slack subtracted.

Cases 4a and 4b illustrate the idea of hub isolation referred to early in the chapter. In case #4a, the network is completely disconnected ($p = 0$) in the sense that each aircraft has all of its stops at only one of the two hubs. The effect is that the scheduled bank times at a hub cannot be disrupted by late arrivals caused by congestion at the other location (they can, of course, be disrupted by earlier delays at the same location). In contrast, case #4b ensures that aircraft encountering delays at one hub have the maximum chance to disrupt the schedule at the second,

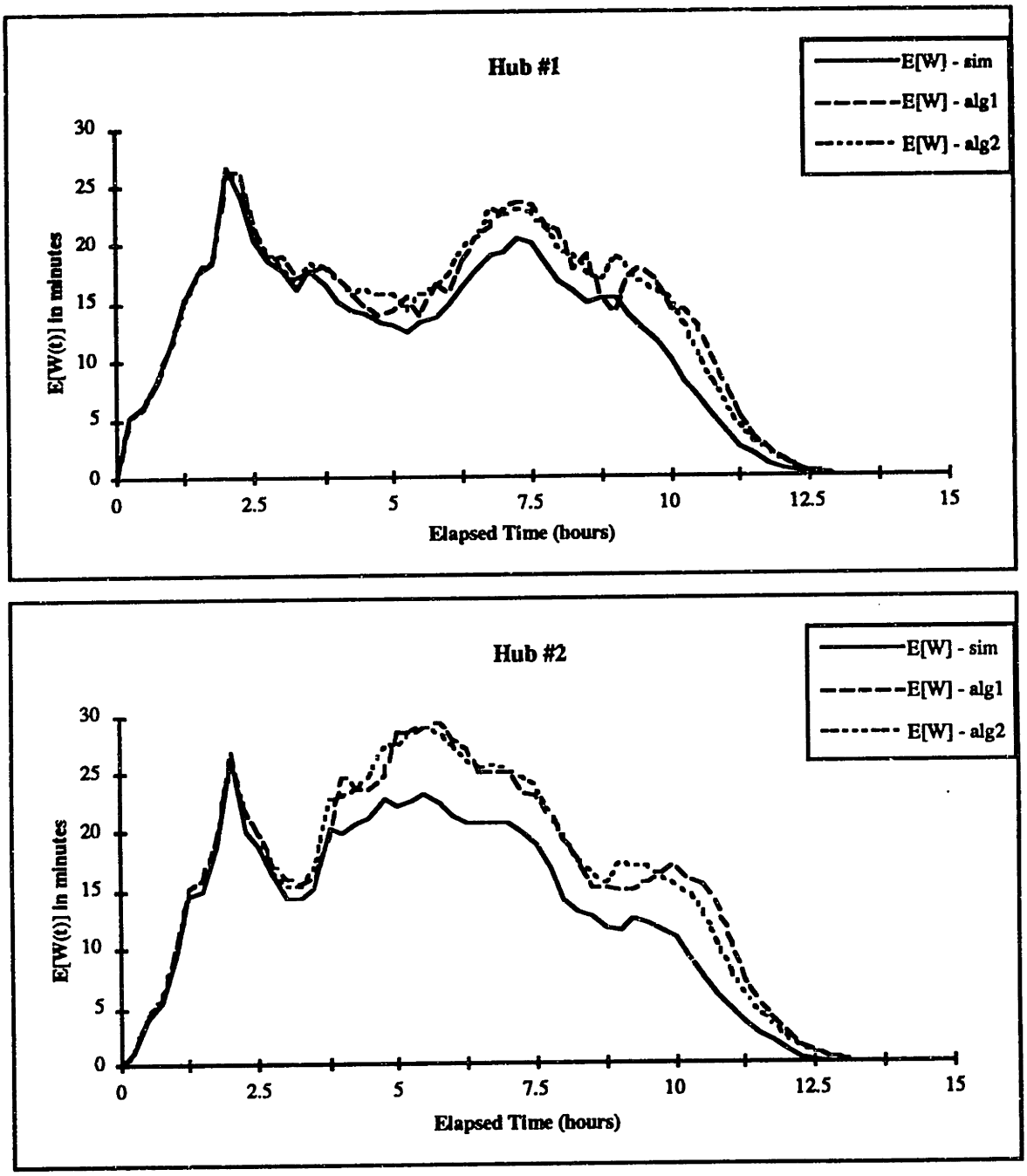


Figure 5.13: Comparison of expected waiting times predicted by simulation and the two decomposition methods for case #3.

since that is their next destination.

In both cases of this experiment, the initial state of the first hub is taken to be 1 (lowest capacity), and the initial state of the second hub is set at 3 (highest capacity). The phenomenon of interest is the propagation of delays created at the first hub to the banks of the second. To examine this, consider Figure 5.14. Note that unlike previous graphs, this figure plots *average cumulative delay per arriving aircraft* rather than the queueing delays present at the hub *when the aircraft arrives*. Thus the early banks[1] show zero delay, while later banks reflect delay carried over from previous points in the itinerary. The figure indicates a degradation in performance at hub #1 when it is isolated, as well as the corresponding improvement at hub #2. Conversely, the fully connected case benefits hub #1 at the expense of #2.

Do these results make sense? Clearly we expect hub #2's schedule to become more reliable when it is disconnected from the disruptions produced by #1. But we also see that hub #1's schedule performance improves when it moves in the opposite direction — from disconnected to fully connected. Examining the situation at hub #1 more closely, we notice that the delays in the connected case seems to lag the delays in the disconnected case by about 2 banks (2 hours). This is no coincidence: in the connected case, the minimum time between any aircraft's successive visit to the same hub is 4 hours (4 banks), while in the disconnected case it is only 2. Thus the schedule delays produced by the congestion at hub #1 are felt 2 hours later at that hub in the connected case, producing the 2-hour lag. However, this lag does not fully explain the difference in the heights of the two curves. In the connected case, late aircraft leaving hub #1 have the opportunity of recovering some of the delay through slack at their next stop (uncongested hub #2). This opportunity is not available in the disconnected case, since the next stop is (congested) hub #1, a fact which explains why the corresponding curve is higher even after we take account of

[1]The x-axis of Figures 5.14 and 5.15 is in terms of *banks* rather than continuous time — thus 2a indicates the first half of the second bank, 7b indicates the second half of the seventh bank, etc.

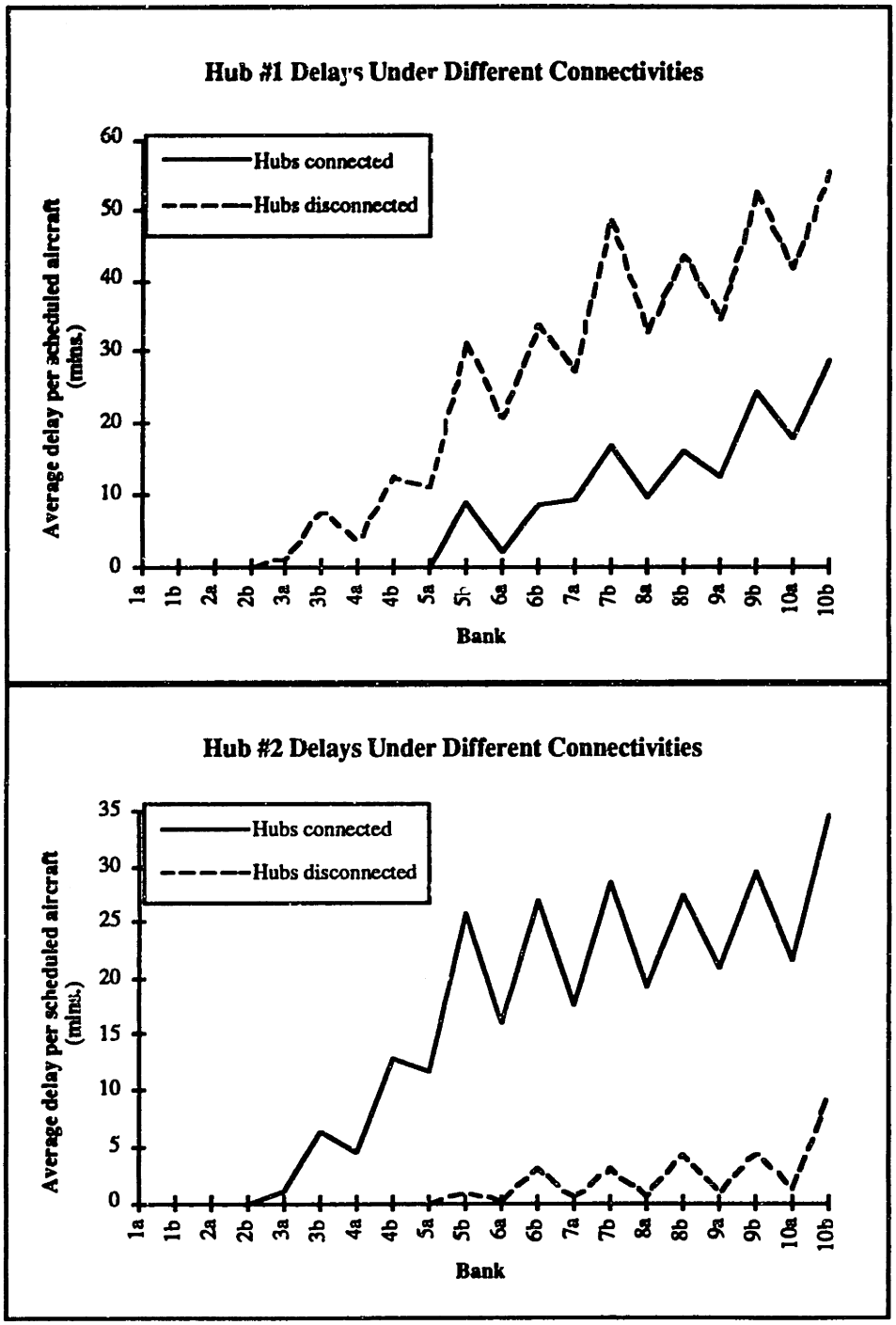


Figure 5.14: Average aircraft delays at two hubs under different degrees of connectivity

the lag.

Cases 5a, 5b, 5c, and 5d illustrate the effect of slack on aircraft lateness. We noted earlier that higher slack preserves demand peaking and may actually increase queueing delays at the hubs. On the other hand, slack reduces each aircraft's *cumulative delay.* Figure 5.15 illustrates that this second effect predominates in this relatively light traffic. For varying slack values, the figure plots the average cumulative delay per aircraft arriving at each bank of the day, not including any waiting at the current stop. Certainly the figure does not contain any surprises. We include it in order to illustrate the kind of planning for which the models are well-suited.

5.3 Concluding Remarks

In this chapter we have developed two related approximation approaches to the difficult problem of modeling transient queueing behavior in a hub-and-spoke network. We would summarize our major findings as follows:

1. *Importance of traffic splitting phenomenon.* High uncertainty in levels of delay encountered by aircraft is a prominent feature of the network problem. Unfortunately, accuracy in keeping track of aircraft amid this uncertainty is limited by high computational complexity.

2. *Continued importance of deterministic effect.* The peaked pattern of demand at hub airports remains a strongly determining factor in predicting waiting times, particularly when major banks are separated by large lengths of time.

3. *Delay and smoothing.* On the other hand, in cases where banks are narrowly spaced, delay propagation exerts a strong smoothing effect on the demand and waiting time profiles.

4. *Effects of hub isolation.* A policy of isolating a congestion-prone hub clearly does have the effect of improving performance at others. On the other hand,

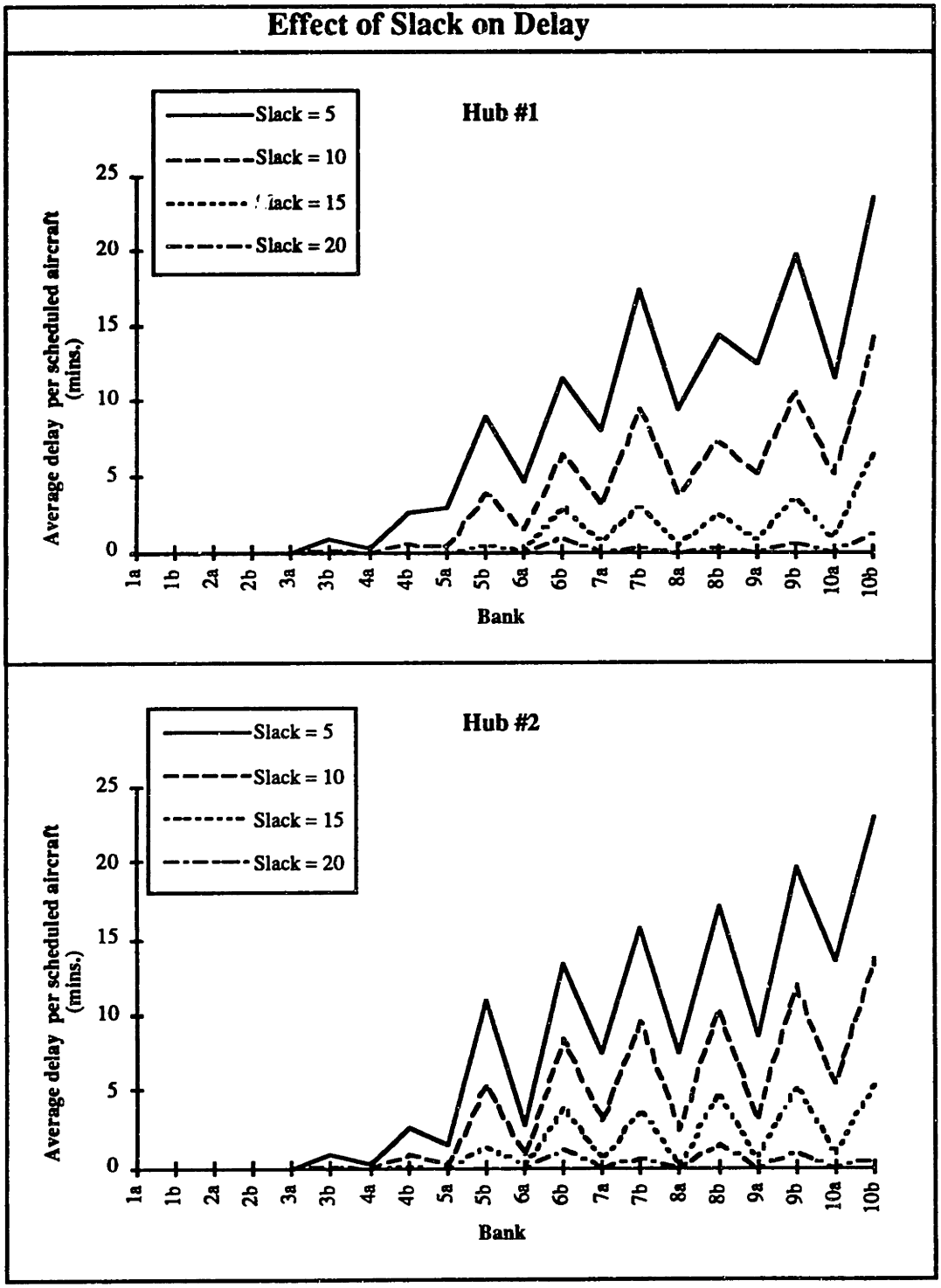


Figure 5.15: Effect of slack on total delay at each hub under 50% connectivity

under this policy the isolated hub produces congestion delays which disrupt its own future schedule.

We conclude with a few remarks about running times. As we reported earlier, the running times for Algorithms 1 and 2 are high even for the small 2-hub test network: approximately one hour for Algorithm 1 and three hours for Algorithm 2 on a DEC-3100 workstation. These times are particular poor considering that the running time for the simulation program (5000 samples) is significantly ***shorter*** — about 10 minutes. In the absence of improvements in the algorithms, this observation favors simulation. However, the implementation of Algorithms 1 and 2 used in our tests is a rather inefficient one. Incorporating the earlier suggestion that the recursion be restarted every m periods rather than at every period would reduce running times by at least a factor of

$$\frac{K+1}{(K/m)+1} \approx m$$

A value $m = 10$ (2 1/2 hours), which is approximately the minimum time a typical aircraft would have between successive visits to hubs, would reduce running times by a factor of 9 (for $K = 80$ periods). This improvement alone would bring the running times of the algorithms into the same range as simulation. The reduction is important for the general problem because the number of simulations necessary cannot be known in advance. However, at least in this test case, the simulation procedures themselves, based on the same ideas of the original Markov and semi-Markov capacity models, offer a third approach to understanding network effects.

Chapter 6

Conclusion

The main contribution of the thesis has been the development of queueing models which are appropriate for congestion at hub airports and in hub-and-spoke systems. Because of the need to describe the transient behavior of such systems, these models depart considerably from traditional queueing theory methods. In this chapter we summarize the main results of the models and indicate possibilities for further research.

6.1 Summary of Main Results

We summarize the principal results of the thesis as follows.

1. *Development of a recursive algorithm for single airport.* The recursive computational scheme presented in Chapter 3 adequately models the principal uncertainty in the landing queue (capacity) while retaining tractability in treating transient behavior. The key to the algorithm's success lies in the division of time into short intervals where computation is simple and the provision of a capacity model which unifies the behavior in these individual intervals into a coherent whole.

2. *Adaptation of diffusion approximation.* In Chapter 3 we also introduced an adaptation of the classical diffusion approximation for a single queue with

correlated service times.

3. *Insights about queue behavior from test runs.* The experimental runs of the recursive model, both in Chapter 3 and in Chapter 4, underscore several important features of this kind of queueing system: high variability, strong memory of starting conditions (due to capacity correlation), and slow approach to steady state.

4. *Study of interaction between airlines at hubs.* The study of the schedule at Dallas-Fort Worth in Chapter 4 suggests that when there is more than one principal carrier at a hub, schedule positioning plays a role in allocating queueing delay. Our analysis in particular implies that Delta would benefit by scheduling greater slack between itself and preceding American Airlines banks at the busiest times of day.

5. *Effects of demand smoothing policies.* In Chapter 4 we also studied the relationship between the smoothing of demand and delays. Our results indicate that small amounts of traffic smoothing at DFW may lead to substantial reductions in day-to-day congestion delay, while further smoothing of traffic exhibits rapidly diminishing benefits and increasing costs.

6. *Evaluation of decomposition algorithms for networks.* In Chapter 5 we developed and tested two queueing algorithms which take into account the network effects at hub airports. We showed that these decomposition approaches work fairly well, although tests against a simulation procedure show them to underestimate the true spreading of demand which takes place because of delays.

7. *Study of hub-and-spoke network queueing effects.* In airline networks, delays can have a smoothing effect on demand, which in heavy traffic situations may actually work to alleviate higher delays. This effect may be important at airports which have extended periods of heavy traffic or narrowly spaced

banks. At DFW, ample separation between the highest banks makes it less of an issue.

8. *Examination of hub isolation policies.* At the close of Chapter 5 we addressed the issue of hub isolation as a strategy for reducing the disruptions in the network caused by delays at a chronically congested hub. Our results suggest that by isolating the problem hub, a carrier eliminates schedule disruption at the other hubs in the system, but the cost may be a worsened performance at the congested hub itself.

6.2 Directions for Further Research

Because much of the work we have undertaken here represents a new approach to a difficult problem, there are numerous ways in which it can be extended. These are grouped here into four categories: model validation issues, queueing theoretic questions, improvements in implementation, and new applications.

1. *Model validation.* As we have indicated in Chapters 4 and 5, our validation procedures for both the single-queue and network models are incomplete, mainly because of the unavailability of appropriate data. The best way to calibrate the model and obtain a clearer measure of its performance would be to implement it in practice. The same holds for the network models, where the degree of approximation is greater still.

2. *Queueing theoretic issues.* The discussions of Chapters 3 and 5 raise several interesting questions. One concerns the slow convergence to steady state for the single queue model. Odoni and Roth [29] have indicated such slow convergence in systems where service times are purely independent. In our model, service times within intervals are exactly equal, while across intervals they have positive correlation. We conjecture that these characteristics further slow the approach to steady state, but obtaining more definitive analytical results to

prove or disprove this conjecture is an open problem. A second theoretical issue concerns the extension of the diffusion approximation to time-varying demand rates. With such an extension, the method would become a serious alternative for the network problem, where its faster running time would be a distinct advantage.

3. *Improvements in implementation.* There are numerous implementation improvements which can be made in the one-hub and network models. First of all, we note that much of our analysis has focused on *moments*, especially *expected values.* Because queue length and waiting time variances are high, however, our focus could overlook important variability phenomena. Further experimentation with approximate *distributions* is warranted. With respect to the network algorithms, there is considerable room for improvement. Algorithm 2 in particular has considerable flexibility which could and should be exploited. Further testing of the procedures for a more widely varying set of networks and schedules is also warranted.

4. *Further application areas.* Finally, there is a need for further research on applications. Within air transportation, an issue of substantial current interest is that of planning ground holds to match demand with forecasted capacities. There have been a number of static and dynamic optimization schemes proposed, but none of these explicitly accounts for the queueing effects. Incorporating the congestion models of this work is one possible further use of the models we have developed here. Another potential application is in manufacturing. While the majority of models in the literature deal only with steady-state analysis of manufacturing queues, our method offers a way to track transient behavior. Of particular interest are systems where service times are dependent on an external stochastic phenomenon (e.g. machine breakdown). The exploration of such modeling possibilities is clearly an important and relevant task for future research.

Bibliography

[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] AIRBORNE INSTRUMENT LABORATORIES, INC. *Operational Evaluation of Airport Runway Design and Capacity,* Report 7601-6, January 1963.

[3] AIRBORNE INSTRUMENT LABORATORIES, INC. *Airport Capacity: a Handbook for Analyzing Airport Design and to Determine Practical Movement Rates and Aircraft Operating Costs,* June,1963.

[4] G. ANDREATTA AND G. ROMANIN-JACUR. "Aircraft Flow Management Under Congestion," *Transportation Science* **21**:4, 249-53 (1987).

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

[6] ARNOLD BARNETT, TODD CURTIS, JESSE GORANSON, AND ANDREW PATRICK. "Better Than Ever: Nonstop Jet Service in an Era of Hubs and Spokes," *Sloan Management Review* **33**:2 (1992).

[7] 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).

[8] 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).

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

[10] THOMAS COOK, President, American Airlines Decision Technologies. Talk delivered at Operations Research Center, Massachusetts Insitute of Technology, November 15, 1990.

[11] UNITED STATES GENERAL ACCOUNTING OFFICE. *Airline Competition: Higher Fares and Reduced Competition at Concentrated Airports,* Report to Congressional Requesters, July 1990.

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

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

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

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

[16] J.M. HAMMERSLEY AND D.C. HANDSCOMB. *Monte Carlo Methods,* Methuen, London, 1964.

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

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

[19] 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).

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

[21] 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).

[22] 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).

[23] F.P. KELLY. *Reversibility and Stochastic Networks,* John Wiley and Sons, Chichester (U.K.)., 1979.

[24] 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).

[25] 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).

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

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

[28] AMEDEO R. ODONI. "The Flow Management Problem in Air Traffic Control," p. 269-288 in *Flow Control of Congested Networks,* A.R. Odoni, L. Bianco, and G. Szego (eds.), Springer-Verlag, New York, 1987.

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

[30] PEAT, MARWICK, AND MITCHELL, INC. *Techniques for Determining Airport Airside Capacity and Delay,* Report No. FAA-RD-74-124 prepared for the Federal Aviation Administration, June 1976.

[31] PEAT, MARWICK, AND MITCHELL, INC. *Technical Report on Airport Capacity and Delay Studies,* Report No. FAA-RD-76-153 prepared for the Federal Aviation Administration, June 1976.

[32] OCTAVIO RICHETTA. *Ground Holding Strategies for Air Traffic Control Under Uncertainty,* Operations Research Center Technical Report No. 198, Massacusetts Institute of Technology, Cambridge, MA, 1991.

[33] 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.

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

[35] M. TERRAB. *Ground Holding Strategies in Air Traffic Control,* Operations Research Center Technical Report No. 196, Massacusetts Institute of Technology, Cambridge, MA, 1990.

[36] REX S. TOH AND RICHARD G. HIGGINS. "The Impact of Hub and Spoke Network Centralization and Route Monopoly on Domestic Airline Profitability," *Transportation Journal* **24**:4, 249-53 (1987).

[37] PETER B. VRANAS, DIMITRIS J. BERTSIMAS, AND AMEDEO R. ODONI. *The Multi-airport Ground-holding Problem in Air Traffic Control,* Operations Research Center Working Paper No. OR263-92, Massacusetts Institute of Technology, Cambridge, MA, 1992.

[38] PETER B. VRANAS, DIMITRIS J. BERTSIMAS, AND AMEDEO R. ODONI. *Dynamic Ground-holding Policies for a Network of Airports,* Operations Research Center Working Paper No. OR265-92, Massacusetts Institute of Technology, Cambridge, MA, 1992.

[39] W. WHITT. "The Queuing Network Analyzer," *Bell System Technical Journal* 62:9, 2779-2815 (1983).