Dynamic Statistical Models for the Prediction of Aircraft Take-off Times

by

Robert Arthur Shumsky

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1995

Author .....
Department of Electrical Engineering and Computer Science
June 16, 1995

Certified by .....
Arnold I. Barnett
Professor of Operations Research and Management
Thesis Supervisor

Certified by .....
Amedeo R. Odoni
Professor, Departments of Aeronautics and Astronautics and of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Thomas L. Magnanti
Co-Director, Operations Research Center, and George Eastman Professor of Management Science, Sloan School of Management

Dynamic Statistical Models for the Prediction of Aircraft Take-off Times

by

Robert Arthur Shumsky

Submitted to the Department of Electrical Engineering and Computer Science
on June 16, 1995, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Safe and efficient management of air traffic requires accurate predictions of aircraft trajectories. In the existing air traffic system, predictions of take-off times are a major source of forecast error. This thesis presents three primary contributions: forecasting models for predicting individual take-off times, models of aircraft flow to predict departure congestion at major airports, and methods for balancing the frequency of forecast updates with the costs of forecast inaccuracy. The models use real-time data to update parameters and generate forecasts. They are tested with data from the existing air traffic management system as well as data that may be obtained from other sources, such as the commercial air carriers.

Empirical tests using data specific to Logan Airport suggest that the take-off time forecasting models achieve small improvements in forecast accuracy over existing methods when using information currently available to the air traffic management system. If additional, carrier-specific information is available, the models achieve more substantial improvements. Empirical tests also demonstrate that the departure flow models produce accurate predictions of airfield congestion over both ten-minute and one-hour forecast horizons.

Once a take-off time forecast is produced, there are frequent opportunities to update the prediction with new information. We develop forecasting algorithms which balance prediction accuracy and the cost of each forecast update, for frequent updates burden the traffic management system's computers and lead to distrust in a system that cannot 'make up its mind.' Numerical examples demonstrate that the proposed algorithms significantly improve forecast accuracy and require fewer updates than existing procedures.

Thesis Supervisor: Arnold I. Barnett
Title: Professor of Operations Research and Management,
Sloan School of Management

Thesis Supervisor: Amedeo R. Odoni
Title: Professor, Departments of Aeronautics and Astronautics and
of Civil and Environmental Engineering

Acknowledgments

I would like to thank my advisors, Professors Arnold Barnett and Amedeo Odoni. They have been invaluable as collaborators, teachers, and mentors. When initiating this research I benefited from their ability to see through the details to the central questions. Throughout the project they have challenged me to think creatively and write clearly. They have been generous with their time and supportive of my best efforts.

I am also grateful to the readers of this thesis, Professor John D.C. Little and Dr. Eric Blair of the MITRE Corporation. Their well-chosen comments frequently led to fruitful areas of research. I thank Professor Alvin Drake for his advice and encouragement, which were always offered with warmth and good humor. Others who had an impact on the contents of this thesis were Peter Stynes, Ken Howard and Eugene Gilbo of the Volpe National Transportation Systems Center, Richard Italiano of American Airlines, and Michael White and Lee Brown of MITRE. All helped me to gather data and to understand more fully the air traffic management system and air carrier operations.

I appreciate the friendship of many of my classmates at the Operations Research Center, and I owe special thanks to: Ármann Ingólfsson for his unflagging curiosity (and unflagging pace when we ran together); Edieal Pinker for challenging conversation and last-minute proofreading; Kerry Malone for heartening encouragement; Eriko Kitazawa for her enthusiastic cheerleading; Elaine Chew for refreshing my sense of humor; and Alan Rimm-Kaufman for his contagious excitement. Thanks as well to David Markowitz for volunteering his time to care for our workstations. I also appreciate the excellent work of the ORC staff, Paulette Mosley, Laura Rose and Cheryl Mottley, who make the ORC run so smoothly.

My time at MIT has been enriched by the presence of my housemate, Andrea Page. I also have been surrounded by a supportive family, including brother and sister-in-law David and Emily Shumsky and grandmother Beatrice Shambad, who still wonders why they paid me to do this. Most of all, I wish to thank my parents, Jack and Evelyn Shumsky, for their constant love and support.

This project was funded by the Department of Transportation and Federal Aviation Administration under grant 94-G-030. I am grateful for their financial assistance.

Contents

Chapter 1

Introduction

During the past few decades, commercial air traffic has increased dramatically, while the capacity of the system which serves that traffic has not kept pace. The resulting congestion produces costly delays. The Federal Aviation Administration (FAA) estimates that delays initiated by air traffic control cost airlines $1.5 billion per year in direct operating costs in addition to the cost of delays borne by passengers [15]. The FAA also anticipates substantial increases in demand on the air traffic system over the next decade. Expanding the physical capacity of the system might alleviate congestion, but few infrastructure improvements, such as new airports or runways at existing airports, are planned in the most congested parts of the system [25]. One of the FAA's primary research initiatives in response to increasing congestion is to create a system for the strategic management of traffic flow. By controlling the movements of airborne aircraft and by holding aircraft on the ground, traffic managers attempt to reduce congestion and increase system throughput.

Even with perfectly accurate projections of demand and capacity, traffic management is a difficult task. Capacity reduction at a single airport may lead to changes in hundreds of flight plans. The task is made even more difficult by the large uncertainties associated with forecasts of both demand and capacity. There are delays between air traffic management actions and effects, so potential problems must be recognized and corrected far in advance. Effective action depends on accurate forecasts.

To provide up-to-date information to air traffic managers, the FAA has developed the Enhanced Traffic Management System (ETMS).[1] The ETMS consolidates and distributes

[1]Strictly speaking, there are two closely related traffic management technology programs within the

real-time information about the national airspace and can indicate the locations of all controlled aircraft in the U.S. on a single video screen. Recently, a predictive function has been added to the ETMS so that it generates predictions of a flight's path. However, forecasts of aircraft take-off times have a large variance [20].

The following example demonstrates why take-off time uncertainty is a significant problem for air traffic management. Imagine that a large number of aircraft are planning to fly through the same sector (a region of airspace) within one hour. An air traffic manager using the ETMS is concerned that many of the aircraft will reach the sector at the same time, and that the airspace will become dangerously crowded. In addition, suppose that many of these aircraft are still on the ground, either about to push-back from the gate or taxiing to the departure runway. If the traffic manager anticipates trouble, then it is possible to hold some of these aircraft on the ground. The manager, however, faces a decision under great uncertainty. The standard deviation of ETMS take-off time forecasts is between ten and fifteen minutes, and it takes only five minutes to cross the potentially congested sector. Since there are no reliable predictions of when the aircraft will leave the ground, there are no reliable predictions of when they will enter and leave the sector.

The manager will not be sure whether the sector will be 'loaded' until many of the aircraft are in the air. Once airborne, aircraft trajectories are almost deterministic and sector arrival forecasts are accurate, but by then it is too late to keep many of the aircraft on the ground. In general, when take-off time forecasts are inaccurate, air traffic managers may be surprised by the appearance of flights in the sectors near the terminal airspace. They may also take action to prevent projected congestion which may be merely a statistical illusion. Inaccurate long-term forecasts decrease the effectiveness of strategic traffic management actions.

This thesis develops models and forecasting procedures which improve take-off time forecast accuracy. It presents probabilistic models of the aircraft departure process from major U.S. airports. It develops procedures which use real-time data to update these models. The models are used to generate forecasts of aircraft departure demand, departure queue sizes, and take-off times of individual aircraft. Empirical tests investigate the value of the

FAA. The Advanced Traffic Management System (ATMS) is a system for research and development while the ETMS is the version brought into the field. We will use the acronym 'ETMS' when speaking of either program.

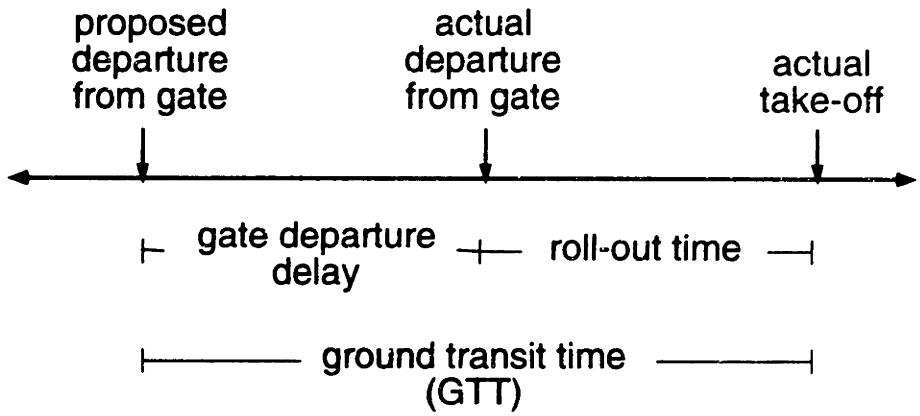


Figure 1-1: Definition of ground transit time

models as tools for monitoring and forecasting airport congestion and aircraft delays.

Section 1.1 provides an overview of the statistical models and results of the thesis. The section specifies the data requirements of the models, describes model form and function, and summarizes the results of empirical tests of each model. Section 1.2 describes the primary contributions of this thesis. Section 1.3 concludes this chapter with a guide to the organization of the thesis.

1.1 Overview of the Models and Summary of Results

Predictions of take-off time are derived from predictions of *ground transit time* (GTT), the time between the proposed departure time from the gate and the actual take-off:

$$\text{actual take-off time} = \text{proposed departure time from gate} + \text{GTT}$$

(see figure 1-1). The *proposed departure time* is the departure time submitted by the airline to the FAA as part of the flight plan. A flight plan may be 'filed' as late as thirty minutes in advance of the proposed departure time. However, many carriers automatically file flight plans weeks in advance, and for most flights by the major carriers, the proposed departure time is identical to the scheduled departure time published for consumers. Therefore, we will assume that for all flights the proposed departure time is known at the time a forecast is produced.

Ground transit times are divided into two components: *gate delay* and *roll-out time.* Gate departure delay is the time between proposed push-back (departure) and actual push-back from the gate. While the average gate departure delay is small, there are occasional

large delays of over one hour. These long gate departure delays have a variety of causes, including mechanical problems or delays to incoming flights. Roll-out time is the time between push-back and actual take-off. It includes time spent turning on the apron near the gate, taxiing to the runway, waiting in a departure queue, and rolling down the runway. Note that a significant portion of roll-out time may be spent waiting, rather than rolling.

The ability to predict variations in gate delay and roll-out time depends on both the look-ahead time, or forecast horizon, of the prediction as well as the information that is available to generate the prediction. Our models will be shaped by the questions: "what do we know, and when do we know it?" Section 1.1.1 provides some answers to these questions by describing the forecast horizon and data requirements of the models. Section 1.1.2 describes the models themselves and summarizes the empirical results.

1.1.1 Forecast Horizons and Data Availability

The design of each model is shaped by both the availability of data and the forecasting requirements it is intended to fulfill. A model which generates forecasts minutes in advance will employ different techniques and data than models with forecast horizons of days or months. We will assume that our forecasts will be used as input to strategic traffic management decisions. At the earliest, these decisions are taken from three to six hours in advance of projected congestion. At the latest, the decision may be taken while an aircraft is rolling out. This range describes our forecast horizon. It is possible that one set of models are most accurate for long-term forecasts while other models are more appropriate over short forecast horizons.

Assumptions about data availability will also shape the models. We might build models under that assumption that all possible data sources are available. For example, the FAA is considering adopting the Global Positioning System (GPS) for aircraft navigation, precision approach procedures for landing aircraft, and airport surface monitors for tracking aircraft on the ground [16]. When linked to a GPS-based surface monitoring system, the ETMS would be able to report the location of each aircraft on the airfield. A detailed model which incorporates this information should produce extremely accurate roll-out time predictions over short time horizons. However, such a system will not be available until the rather distant future. According to the FAA strategic plan, a reliable GPS-based system for terminal airspace navigation will not be available until the year 2000, GPS-based landing

	Currently On-line in ETMS	Potential Addition to ETMS
Environmental Data	weather conditions	runway configuration
Flight Data	proposed push-back time actual take-off time carrier flight duration size of aircraft	actual push-back time
Connection Information	arrival delay	flight connections linking arrivals and departures

Table 1.1: Data sources assumed to be available to the forecasting models

systems are scheduled for 2005, and an airport surface monitor has been proposed but no schedule for development or implementation has been published.

Conversely, if we restrict ourselves to real-time data that is currently available to the ETMS, we lose the opportunity to explore whether additional sources of information may significantly improve forecast accuracy. The approach taken here falls between the extremes described above. We assume that our models will have access to real-time data which is either available now or may be channeled to the ETMS from existing databases. In particular, the models utilize data that are collected in real-time by the ETMS today or that satisfy the following three requirements:

- The data may significantly improve the model's forecast accuracy;
- The data may be obtained for the ETMS, in real-time, using data collection systems that have already been developed by the FAA or the carriers;
- Some subset of the data can be obtained from historical sources.

The third requirement ensures that the models can be thoroughly tested with historical data without relying on simulation. Table 1.1 lists the data that satisfy these three requirements. The data listed in the right-hand column are not currently available to the ETMS.

There may be important sources of information which are not available to us but might be made available to the ETMS. For example, an individual carrier may be aware that a flight must be held for a crew which is arriving on another, late flight. This event, and many others, would appear as statistical noise in the model. One implication of the empirical results of this thesis is that such delay information, if shared by the carriers, may produce greater gains in forecast accuracy than the most complex mathematical models which rely on existing information sources.

1.1.2 Description of Individual Models and Results

In this section we provide an overview of the models and results of the thesis. We first describe relationships between the models and then provide more details about each model.

A primary benefit of improved take-off time forecasts is an improvement in demand forecasts at fixes (an airspace reference point), sectors and airports downstream. The first model developed in this thesis (in Chapter 2) provides motivation for subsequent models of aircraft departures by quantifying the relationship between errors in individual aircraft trajectories and errors in arrival demand predictions at airports. The model demonstrates that current levels of uncertainty in take-off time forecasts are likely to cause significant errors in demand forecasts.

The remainder of the thesis develops models for take-off time prediction. Figure 1-2 displays the relationships among the primary models. Ground transit time forecasts combine forecasts generated by a *gate departure delay model* and a *roll-out time model.* Analysis of data from Logan Airport will demonstrate that at Logan gate departure delay and roll-out time are essentially statistically independent, and therefore the models for each were developed separately.

One would expect that roll-out times would be heavily influenced by congestion on the airfield (any airline passenger who has been told by the pilot that their aircraft is "eighth in line for take-off" would recognize this). Therefore, we constructed an *aggregate flow model* to forecast airfield congestion. While originally developed to provide forecasts of airfield congestion to the roll-out time models, the forecasts of the aggregate flow models should be of independent interest to traffic managers at airports.

The gate delay and roll-out models may be used to generate take-off time predictions until the expected departure time of a flight approaches. Predictions might be revised on

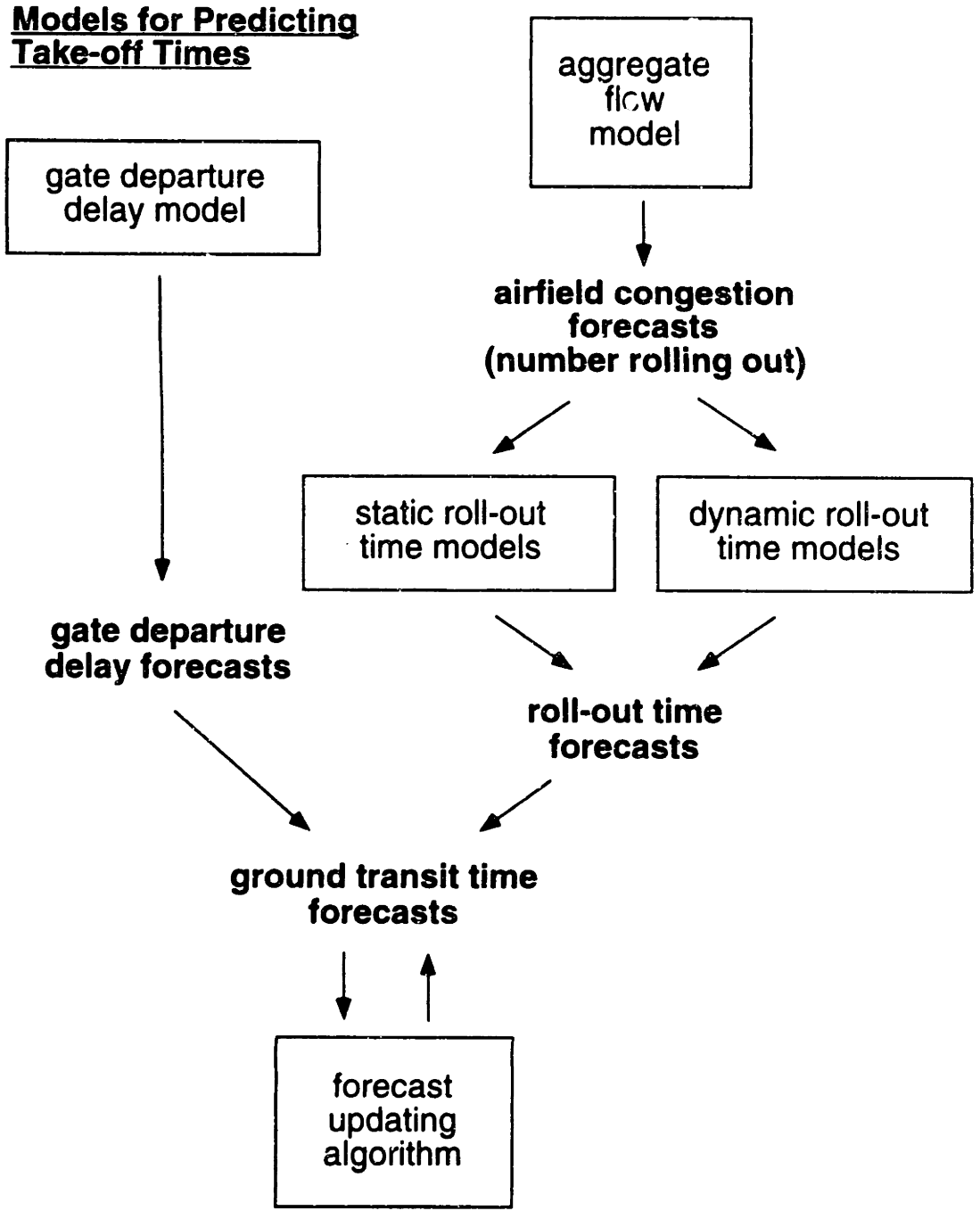


Figure 1-2: Models for gate departure delay, roll-out time, and ground transit time forecasts

the arrival of relevant information, such as an updated weather forecast. When close to the expected take-off time, detailed information about the individual flight's status is available to the ETMS and there is the opportunity to almost continually update forecasts with this information. A *forecast updating algorithm* selects the times to update the forecast and produces new forecasts from the available information.

Following are more details about each model and summaries of the results of empirical experiments using data from Logan Airport.

Gate Departure Delay (Turn) Model

We have found that a simple model linking arrival delays with subsequent gate departure delays produces a 50% reduction in mean squared forecast error from existing methods in a sample of flights from Logan Airport. This 'turn model' simply adds a minute of gate departure delay for every minute of arrival delay when the time between gate arrival and scheduled gate departure falls below some threshold. The threshold varies among carriers, and we find thresholds for individual carriers by data analysis.

We find that additional delays are correlated with poor weather and other environmental factors. However, models constructed from these factors only obtain marginal improvements in forecast accuracy. The airlines themselves have information that should allow for significantly more accurate predictions of gate departure delay. The thesis describes the turn model and then discusses the potential benefits of increased communication between the FAA and the carriers.

Aggregate Flow Model

In order to produce real-time forecasts of departure service rates and airfield congestion, we developed aggregate flow models which monitor and predict the number of aircraft rolling out. The models collect aggregate statistics on departures in discrete time periods (15 min.) in order to smooth over small fluctuations in service rates. Simple difference equations describe the movement of aircraft onto and off of the airfield, while real-time data are used to update estimates of the service rates and queues. Empirical tests demonstrate that the models produce accurate forecasts of the number of aircraft on the airfield. The aggregate flow models show promise in monitoring and forecasting airport congestion.

Roll-out Models

Forecasts of roll-out time are generated by a variety of models: an exponential smoother, a lookup table (the current ETMS method), and static and dynamic linear models. A static model gives equal weight to all previous observations, and its parameters are not sensitive to temporary variations in observed roll-out times. A dynamic model gives more weight to recent observations, and should respond more readily to temporary changes in conditions. An important point is that out 'static' models do vary their forecasts among flights. Each forecast produced by a static model is tailored to a flight's operating characteristics, such as the carrier, runway configuration, weather, and a measure of airport congestion. The model is static in the sense that the presumed impact of each of these conditions remains constant over all flights.

The static and dynamic linear models may also be divided into two sub-groups. The first sub-group generates its forecasts from the congestion forecasts produced by the aggregate flow models as well as other factors, such as weather conditions. The second sub-group does not rely on congestion forecasts and replaces them with a rough estimate of departure demand based on the the carriers' scheduled departure times.

This second sub-group produced the most accurate models over long forecast horizons while the linear models with congestion forecasts produced the most accurate short-term forecasts. This is not surprising, for the congestion forecasts from the aggregate flow model become less reliable as the forecast horizon increases. Given congestion forecasts, the dynamic models did not improve over the static models, possibly because the congestion forecasts themselves incorporated the available information about temporary airfield conditions. None of the models were successful in predicting the largest deviations in roll-out times.

Forecast Updating Algorithms

The models developed in this thesis produce predictions of aircraft take-off times. In the traditional forecasting problem, an optimal prediction minimizes an expected forecast error cost under the assumption that the forecast will not be revised. This criterion for optimality ignores the dynamic nature of ETMS forecasts, for an initial prediction may be updated with information that is collected as the expected take-off time approaches. The number of

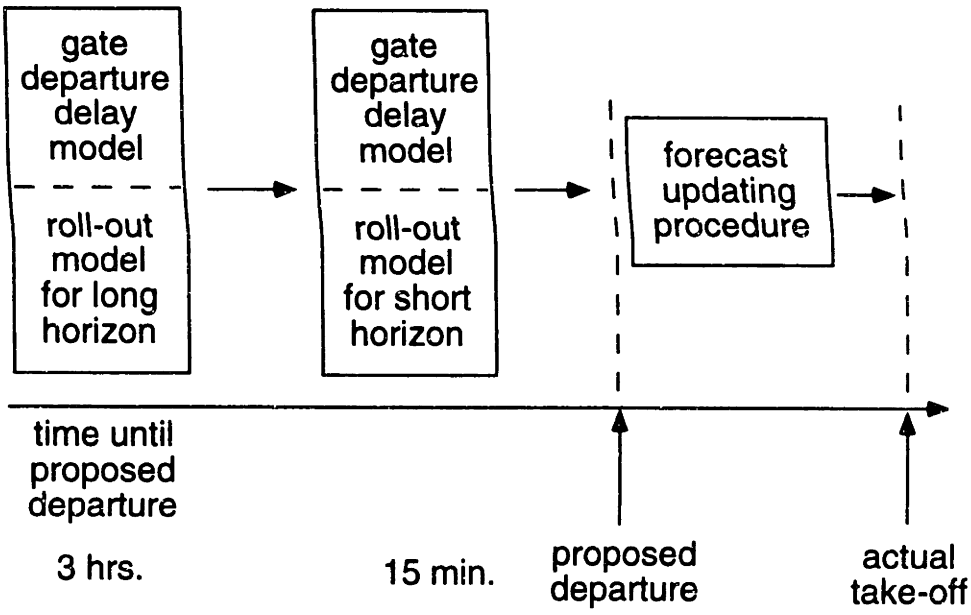


Figure 1-3: Evolution of forecasting methods as flight departure approaches

these updates may be limited by the cost of collecting, processing and distributing relevant information. We generalize the forecasting problem by proposing a cost function for a sequence of forecasts. The cost includes the cost of forecast errors made at various times before take-off as well as an update cost for each forecast revision. A method is derived for finding an optimal sequence of forecasts, given a particular update schedule. When there is a finite number of potential update epochs, a dynamic programming algorithm finds an optimal update schedule, given a sequence of forecasts. In principle, one could iterate this procedure, but we find that after performing these two steps the resulting solution has a forecast cost which is within a few percentage points of a lower bound. This lower bound is achieved when the forecast is continuously updated.

When these algorithms are applied to examples of ETMS take-off time forecasts with high variance, we find that forecast accuracy can be significantly improved by the optimization procedure. In contrast, increasing the number of updates allowed in the examples does relatively little to reduce forecast error. When applied to more typical forecasts with a smaller variance, the advantages of the optimal update schemes are marginal, and simple heuristics are close in performance.

Figure 1-3 displays the transition from model to model as the proposed departure time approaches and the forecast horizon decreases. For a forecast horizon of a few hours, a gate departure delay model and a roll-out time model produce take-off time forecasts. As the

proposed push-back time approaches, GTT forecasts are derived from the gate departure delay model and a roll-out time model that is more adaptable to local conditions (i.e. which includes estimates of temporary airfield congestion). Finally, when the actual push-back is observed, the gate departure delay model is abandoned (the plane is no longer at the gate) and the roll-out time model produces a final roll-out time prediction, along with information about the distribution of forecast error. Thereafter, forecasts are revised according to an update schedule and forecasts derived from the shape of the error distribution.

1.2 Primary Contributions of the Thesis

The main goal of this thesis is to accurately forecast the take-off times of flights from major airports. To this end, we implemented a variety of models, and these led to the following contributions:

- *Forecasting models for aircraft gate delays and roll-out times.* When using information currently available to the air traffic management system, these models achieve small improvements in forecast accuracy over current practice. When additional information not currently available to the system is added, such as connections between arriving and departing flights, the forecasting improvements are substantial.

- *Aggregate flow models to represent the dynamics of departing airport traffic.* These models require information which is not yet available to the ETMS (gate push-back of departing flights), but the success of these simple models in predicting airport congestion suggests the value of this information.

- *Formulation and optimization of a sequential forecasting problem.* In this problem, forecast accuracy is balanced against the cost of forecast revision. We specify a forecast error cost which accumulates over time, and the optimization approach determines a schedule of updates, as well as optimal forecasts, given the schedule. The procedures reduce forecast errors and the number of updates from current ETMS procedures.

Lessons learned from this thesis also have implications for practical implementation in the air traffic management system. The thesis highlights those sources of information that are likely to be most valuable for improving forecasts. Since the FAA is in the process of

upgrading the ETMS to incorporate more data sources, this information may help the FAA to efficiently allocate its resources.

1.3 Contents of the Thesis

Chapters 2 and 3 provide motivation and background for the thesis. Chapter 2 describes the role of forecasting in the air traffic management system and quantifies the level of forecast inaccuracy in the current version of the ETMS. Since traffic management decisions are often driven by forecasts of demand at airports and en-route sectors, Chapter 2 also develops a simple probabilistic model to describe the relationship between errors in trajectory forecasts for individual flights and errors in forecasts of aggregate demand. With the value of accurate forecasts established in Chapter 2, Chapter 3 initiates the analysis of data collected from the ETMS and other sources. In later chapters these data will be employed in empirical tests, so it is important to understand how the data were obtained. The chapter describes the data sampling procedures, specifies the techniques for 'matching' multiple data sets, and provides descriptive statistics of the data. The chapter also tests for sampling bias and describes how this may affect our analysis.

These introductory chapters do not contain a separate literature review, for there is no single body of work which is applicable to the thesis as a whole. Instead, references to relevant material will be made in the introductory sections of appropriate chapters.

Chapters 4, 5, and 6 describe the gate departure delay, aggregate flow, and roll-out models, respectively. In each chapter models are formulated, fitted using historical data, and tested for forecast accuracy. We also test the sensitivity of the models to their parameters, and examine residuals to determine how the models may be improved with additional data.

Chapter 7 describes procedures for updating forecasts once they have been generated by the forecasting models. The chapter describes the cost function developed for this application, which takes into account both the cumulative cost of forecast errors made over time and the cost of updating each forecast. Procedures for finding optimal forecasts and update schedules are derived, and these are tested on examples which are representative of existing air traffic management conditions.

Chapter 8 summarizes the results of the thesis and makes recommendations for implementation. The role of forecasting in air traffic management over short and long time

horizons is examined. The chapter also describes how the aggregate flow models might be incorporated into a real-time airport monitor for air traffic managers. The chapter makes recommendations about which models should be implemented and which information sources would be most valuable for improving ETMS forecasting performance. The thesis ends with suggestions for further research on real-time forecasting models and traffic management.

Chapter 2

The Impact of Forecast Errors on Air Traffic Management

This chapter provides both qualitative and quantitative motivation for improving take-off time forecasts. It first presents a qualitative description of the role of forecasting in air traffic management (ATM) and then presents quantitative evidence of the need for improvements in take-off time forecasting. Section 2.1 describes the relationship between forecasting and air traffic management. The section provides examples of ATM decisions that use as input forecasts of demand at air traffic fixes, sectors or airports. Section 2.2 summarizes empirical studies which determine the size of forecast errors in the current ATM system. The section then presents a probabilistic model which quantifies the relationship between errors in forecasts for individual aircraft and errors in demand forecasts. The model demonstrates that current levels of take-off time forecast errors lead to substantial errors in demand forecasts.

2.1 The Role of Forecasting in Air Traffic Management

While the responsibility of the familiar *air traffic controller* is to maintain separation between individual aircraft, *air traffic managers* are responsible for the strategic and tactical management of traffic flow. ATM personnel are located in the Air Traffic Control System Command Center (ATCSCC) near Washington D.C., the regional Air Route Traffic Control Centers (ARTCCs), the Terminal Radar Approach Control Facilities (TRACONs), and the airport towers. ATM personnel are known by a variety of titles, but we will use here the

term 'air traffic manager.'

The ATM procedure with the largest impact is a Ground Holding Program (GHP) generated by the ATCSCC. When weather is expected to reduce significantly the capacity at major airports, aircraft departing for those airports are assigned departure times that may be minutes, or even hours, after the aircraft's scheduled departure time. The departure times are assigned so that projected operations fall below projected capacity at the affected airports.

Other ATM actions are taken by the regional ARTCCs. If a particular airport expects a moderate amount of congestion, it may request *metering* of aircraft from adjacent air traffic control regions so that aircraft arrive at a pre-arranged rate. When metering has been requested, a local ground holding program may be implemented so that flights to the congested airport from nearby regions are held on the ground. For certain airports metering is in effect almost every day. For example, on weekday afternoons there is a cap on the rate of aircraft from airports in the New England region to Chicago's O'Hare airport. Flights from Logan to O'Hare are routinely held on the ground so that the arrival rate at O'Hare is not excessive.

While there are a variety of other traffic management strategies, all share an important attribute: a lag between action and effect. Large reductions in arrival demand at an airport can be realized by holding aircraft on the ground, but this decision must be made hours in advance of the anticipated congestion. Therefore, the decision to hold flights is based on congestion and demand forecasts rather than on direct observations. The decision may only be as good as the forecasts on which it is based.

The formulation of a Ground Holding Program illustrates the relationship between effective ATM and accurate forecasts. In the current system, GHPs are based on a heuristic which assigns departure times on a FCFS basis [18]. The heuristic assumes that airports have deterministic (known) capacities and that each flight has deterministic (known) departure and travel time. The departure and travel times are used to generate predictions of airport departure and arrival demands. Demand is usually expressed as the number of operations predicted to occur within discrete time periods.

While recent research has led to sophisticated techniques for finding optimal GHPs, proposed algorithms which solve problems of a reasonable size also assume perfect forecasts of capacities and demand [40]. If the forecasts are inaccurate, the 'optimal' ground delay

program may be far from optimal in actuality. Because of the variability in weather, capacity forecasts in particular can be very inaccurate. Even when weather conditions are known, actual capacity can change over time and may vary far from the generally accepted capacity estimates for particular airport conditions. In order to capture these uncertainties, Richetta and Odoni formulate a stochastic program to find an optimal GHP ([29]). Their formulation only applies to a single airport over a small number of weather scenarios, but the uncertainties in capacity projections are likely to motivate further work in stochastic optimization.

Forecasts of demand, rather than capacity, are closer to the concerns of this thesis. The problem solved by Richetta and Odoni also assumes perfect forecasts of take-off and travel times, and therefore assumes perfect forecasts of departure and arrival demand. The same assumption is made in practice when GHPs are formulated at the ATC System Command Center. Prior to the formulation of a GHP, traffic managers obtain demand predictions by counting the number of *scheduled* arrivals to the congested airport. However, even without a GHP, the scheduled arrival pattern will not match the actual arrival pattern.

For example, figure 2-1 compares the actual number of take-offs/hour from Logan airport on March 11, 1991 with a prediction based on the aircrafts' scheduled times of departure. The predictions are calculated by adjusting each flight's scheduled departure time by a constant ground transit time and then counting the number of adjusted take-off times in each hour-long period. The graph shows that the schedule is sometimes an unreliable guide to the actual number of take-offs. In both the morning and afternoon there are large differences between the scheduled and actual demands. One of the periods in the morning has an error of almost twenty flights.

Since GHPs are formulated on the basis of arrival demand, rather than departure demand, this figure does not directly quantify the effect of forecast uncertainty on flow management decisions. However, it demonstrates how individual take-off time forecast errors produce errors in aggregate demand forecasts. These demand forecast errors may have a nonlinear effect on traffic management decisions, with one error having no effect and a slightly larger error leading to an unnecessary control action. It is possible, for example, that flow management actions may be taken unnecessarily to diminish the 'phantom' bulge in demand shown in Figure 2-1 at 17:00. The next section quantifies the extent of these demand errors.

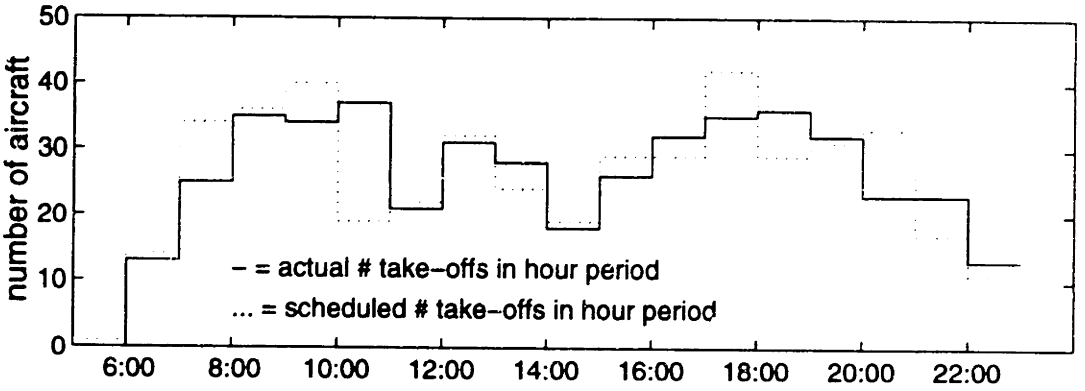


Figure 2-1: Predicted and actual departure demand at Logan Airport during hour-long periods on March 11, 1991. Predictions are based on the scheduled departure times of flights.

2.2 Quantifying Take-off and Demand Forecast Errors

This section quantifies take-off time and demand forecast errors in the existing air traffic management system. Section 2.2.1 describes empirical studies which find large uncertainties associated with take-off time forecasts. Section 2.2.2 presents a simple model which derives errors in demand projections, given a distribution for errors in take-off times.

2.2.1 Empirical Studies of Forecast Accuracy

The most recent published study of overall ETMS take-off time forecast accuracy was produced by ETMS personnel in 1992 [38]. The study averaged forecast errors for flights from thirteen major airports over nine months. The mean difference between actual and forecast take-off time was 14 minutes. In other words, the predictions were *biased* by 14 minutes; on average, aircraft were predicted to depart almost one-quarter of an hour before they actually got off the ground.

Since 1992, ETMS personnel have been developing an improved take-off time prediction algorithm which produces forecasts based on the prevailing weather conditions, time of day, day of week, and other factors [37]. This system is currently being installed in the ETMS, and an experiment by Peter Stynes at the Volpe Transportation Systems Center indicates that the new algorithm does improve forecast accuracy over the previous system. The experiment involved 15,116 flights from O'Hare Airport, and the forecasts produced

	Mean Error (min.)	Mean Absolute Error (min.)	Number of Flights
Bad Weather	6.2	17.0	2747
Good Weather	0.8	9.7	12,369
All Flights	1.7	11.0	15,116

Table 2.1: Forecast accuracy of the most recent ETMS take-off time model, based on approximately 15,000 flights from Chicago O'Hare Airport

by the algorithm had an overall mean error of 1.76 minutes and a mean absolute error (MAE) of 11.0 minutes. The sample was also broken down into 'good weather' and 'bad weather' averages, where bad weather was indicated by standard minimum ceiling heights and visibilities. Table 2.1 displays results from these experiments. This represents an improvement, but there are still large errors in take-off time predictions, especially in bad weather.

While these studies measured the accuracy of ETMS predictions for aircraft on the ground, Goranson compared trajectory forecasts for aircraft on the ground with forecasts for aircraft already in the air [20]. For aircraft traveling through a random sample of 12 sectors (4 each from low, high and super high sectors) he calculated the 'midpoint tardiness,' the difference between the actual and predicted times for the midpoint of the aircraft's visit to the sector. Over all flights, mean midpoint tardiness depended on the forecast horizon, but was always between 8 and 10 minutes. If the sample were limited to flights already in the air, however, the mean midpoint tardiness was always less than one minute. Interviews with air traffic managers confirm the results of this study. These ETMS users agree that forecasts are unreliable before an aircraft takes off, but are much more accurate once the aircraft is off the ground [18] [7].

Goranson also examined the accuracy of the ETMS Monitor Alert, a system designed to warn air traffic managers when the ETMS anticipates congestion in a sector or airport. The study found that the ETMS is a conservative system which calls many false alerts for congestion that does not materialize and allows few surprise alerts for unanticipated congestion. However, Goranson's study was not able to specify the causes of errors in predicted demand. These errors may have been caused by incorrect predictions of sector crossing times, errors in predictions of aircraft trajectories, or surprise cancellations.

No published study has quantified the relationship between individual take-off time

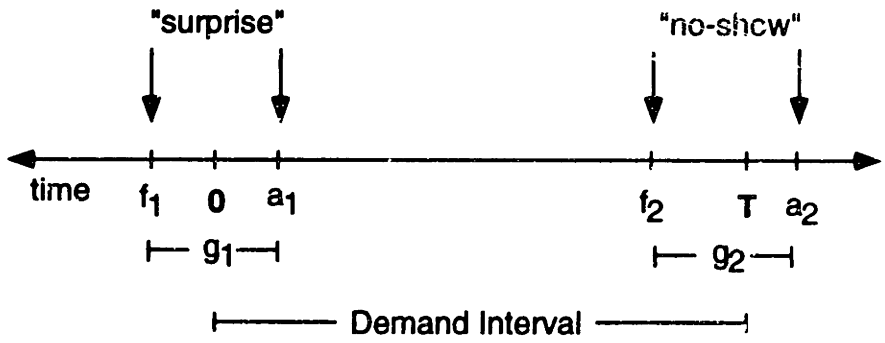


Figure 2-2: Surprises and no-shows during one demand period

forecast accuracy and accuracy in demand forecasts. This is the link between the goal of this thesis (more accurate take-off time forecasts) and many traffic management procedures (which depend upon demand forecasts). This will be the subject of the next section.

2.2.2 A Model of Demand Forecast Error

The research by Goranson and the experiments by Stynes have demonstrated that there are large errors in take-off time forecasts. However, these studies do not quantify the relationship between errors in trajectory forecasts of individual aircraft and errors in aggregate demand forecasts. We ask the question: if this thesis improves take-off time forecasts, to what extent will demand forecasts improve? In order to answer this question, we derive a relationship between forecast error in the arrival times of individual aircraft and forecast errors in arrival demand.

We focus attention on arrival demand since this is an important input when GHPs are formulated. The model described here assumes that the arrival time forecast error is a random variable g with distribution F_g. Results cited in the previous section demonstrate that aircraft trajectories are, in general, quite predictable once the aircraft has left the ground, so F_g will be similar to the distribution of take-off time forecast error. Improvements in the accuracy of take-off-time forecasts should lead to similar improvements in the accuracy of arrival time forecasts.

Formulations for the GHP use the predicted arrival times of aircraft to calculate the demand. Demand is specified as the number of planes predicted to arrive during a discrete time period. Let T be the length of each demand period (T is 15 minutes in the GHPs implemented by the central flow control facility). Figure 2-2 displays a demand period that

begins, by definition, at time 0 and ends at time T. Two pairs of arrivals and forecasts are also shown in the figure. The time a_i is the actual arrival time of aircraft i and f_i is the arrival time forecast for aircraft i. Forecast error g_i is distributed according to F_g. Define a "surprise" to be a flight which is predicted to arrive outside the demand period but instead arrives during the period, while a "no-show" is a flight which is predicted to arrive during the period but does not. Note that a "no-show" eventually does 'show'. The flight arrives during a different demand period, for we assume that there are no unexpected cancellations. Define the following random variables:

$$
\begin{aligned}
l &= \text{actual number of arrivals during the demand period} \\
m &= \text{number of surprises during the demand period} \\
n &= \text{number of no-shows during the demand period} \\
e &= \text{error in predicted demand} \\
&= (l - m + n) - l \\
&= n - m
\end{aligned}
$$

We are interested in the relationship between the distribution of arrival time forecast error (F_g) and the distributions of m, n, and e. Here we derive these distributions for a single period. Demand errors between periods will be heavily correlated, for no-shows and surprises in one period will be mirrored by surprises and no-shows in nearby periods.

For this illustrative model, we assume the following simple probability model for arrivals and forecast errors. The assumptions are first stated here, and then discussed below:

- Actual arrivals occur according to a homogeneous Poisson process with rate λ;
- Arrival time forecast errors for each arrival are independent and are distributed according to the function F_g.

On first glance, both of these assumptions are suspect. Aircraft must maintain a minimum separation, contradicting the assumption of Poisson arrivals. Arrival rates are often time-varying (not homogenous), with sharp peak periods. Forecast errors may be correlated, as when unexpected weather conditions or congestion affect multiple flights.

However, these assumptions are reasonable models for many airports over short periods of time. Major airports have multiple arrival streams into the terminal area, and these

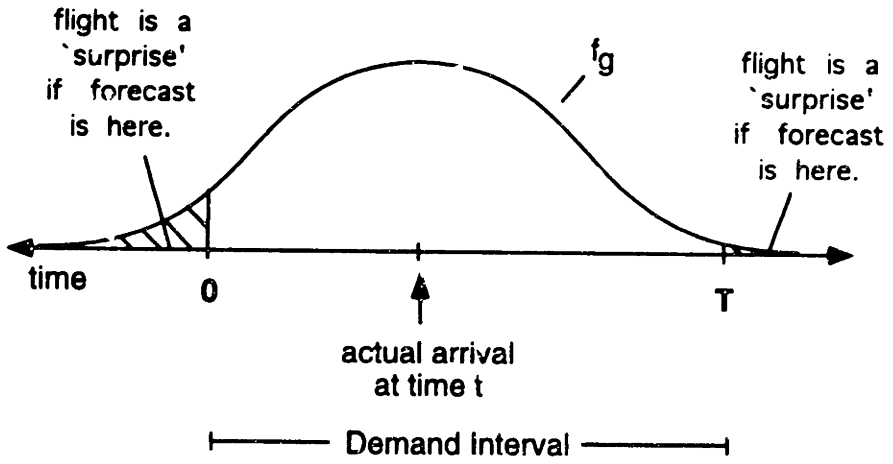


Figure 2-3: A surprise at time t must have a forecast which falls outside of the demand period.

streams operate more or less independently. The sum of the arrival streams may be approximated with a Poisson process. In addition, airports have periods of high or low demand which may last for an hour or more. For the purposes of this illustrative model, we assume that these periods have an approximately homogeneous arrival rate. Ultimately, it would be helpful to test this assumption, but accurate data on arrival times of aircraft to airport terminal areas were not available.

There are two reasonable justifications for the assumption that arrival time forecast errors are independent. Arriving aircraft hail from a variety of airports, and take-off time errors at those airports may be independent. In addition, an effective forecast system would filter out most correlation between departures. If forecast errors were heavily correlated, predictions could be improved by taking the correlation into account.

Given these assumptions, we find the distributions of m and n as well as the mean and variance of e. In fact, both m and n are distributed as Poisson random variables with the same mean, which is a function of F_g, λ, and T.

Number of Surprises (m)

In order for an aircraft to be a surprise, it must arrive within the demand period, but its forecast must predict an arrival outside the demand period. From figure 2-3, any arrival

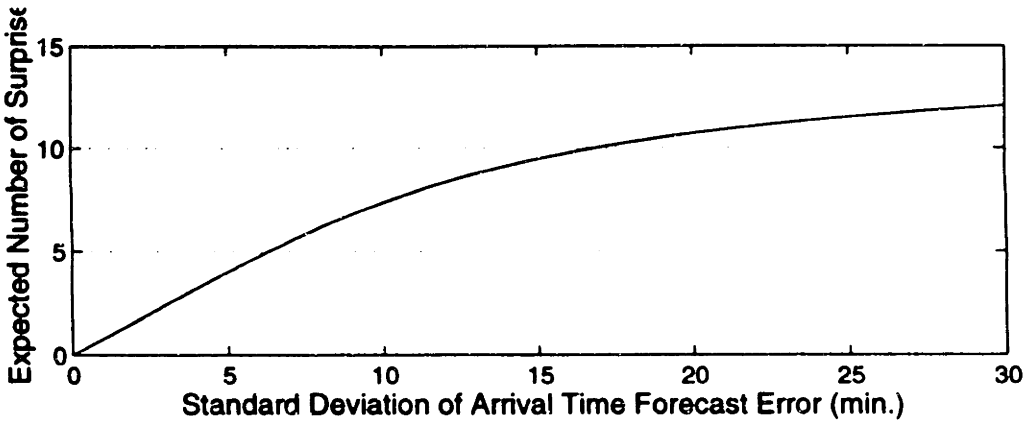


Figure 2-4: Expected number of surprises (λ_m) in the 15-minute period for increasing standard deviation of arrival time forecasts (σ_g). For this figure, $\lambda = 1, g \sim N(0, \sigma_g^2)$, and the expected value has an asymptote at $\lambda T = 15$. The same curve describes the expected number of no-shows.

at time t, $0 \leq t \leq T$, has the following probability of being a surprise:

$$P_m(t) = F_g(-t) + (1 - F_g(T - t)). \tag{2.1}$$

From the definition of a Poisson process, the probability of a surprise arrival in some small interval $[t, t + \Delta]$ is $P_m(t)\lambda\Delta$, independent of all other intervals. The probability of no surprise in the interval is $1 - P_m(t)\lambda\Delta$. Therefore, surprises occur during the demand period according to a *nonstationary* Poisson process with rate $P_m(t)\lambda$. It is well known that m, the total number of surprise arrivals during the period from 0 to T, is distributed as a Poisson random variable with mean:

$$\lambda_m = \int_0^T P_m(t)\lambda dt \tag{2.2}$$

$$= \lambda \int_0^T [F_g(-t) + 1 - F_g(T - t)]dt \tag{2.3}$$

(see [30], p. 46). This expression may be evaluated numerically for an arbitrary distribution F_g. An algebraic expression for λ_m can be derived if forecast errors are normally distributed. The expression is particularly simple if we make the reasonable assumption that forecasts are unbiased, so that $g \sim N(0, \sigma_g^2)$. Let f_g be the density of random variable g. By rearranging and evaluating integrals is equation 2.3, we find that:

$$\lambda_m = \lambda[T(1 - F_g(T) + F_g(-T)) + (I_g(0, T) - (I_g(-T, 0))] \tag{2.4}$$

where

$$I_g(a,b) = \int_a^b t f_g(t) dt \tag{2.5}$$

If the density $f_g(t)$ is symmetric about a mean of zero, then this expression simplifies to:

$$\lambda_m = 2\lambda[T(1 - F_g(T)) + I_g(0,T)] \tag{2.6}$$

When $g \sim N(0, \sigma_g^2)$,

$$I_g(a,b) = \sigma_g^2(f_g(a) - f_g(b)) \tag{2.7}$$

so that

$$\lambda_m = 2\lambda[T(1 - F_g(T)) + \sigma_g^2(f_g(0) - f_g(T))] \tag{2.8}$$

Figure 2-4 displays the mean number of surprises (λ_m), as the standard deviation of the arrival time forecast error (σ_g) increases. The calculations assume that $T = 15$ min. and that $\lambda = 1$ (one aircraft every minute, or 60/hour, on average). The mean λ_m has an asymptote at $\lambda T = 15$, the expected number of arrivals during each fifteen-minute demand period. If forecasts are extremely inaccurate, many of the actual arrivals in the demand period are predicted to arrive during other periods. In the limit, all actual arrivals are surprises, and surprises occur according to the Poisson process with rate λT. To approximate the current level of uncertainty in the air traffic management system, set the arrival-time forecast standard deviation to 10 minutes. If $\sigma_g = 10$, the expected number of surprises (λ_m) is 7.4; on average, 50% of the arrivals during the demand period will be surprises.

Number of No-Shows (n)

By an argument similar to the argument made for surprises, the number of no-shows is also distributed as a Poisson random variable. The probability of a no-show at time t is:

$$P_n(t) = \begin{cases} F_g(t) - F_g(t-T) & \text{if } t < 0 \\ F_g(T-t) - F_g(-t) & \text{if } t > T \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

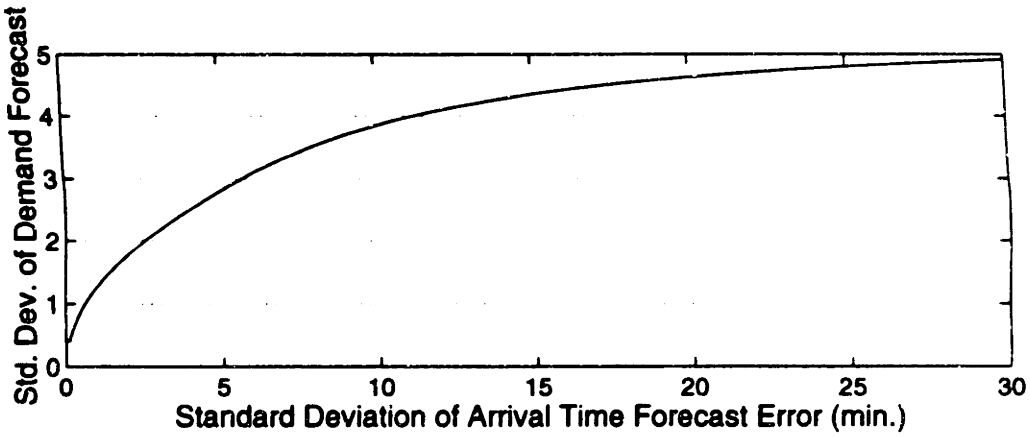


Figure 2-5: Standard deviation of demand forecast errors in 15-Minute period as arrival-time forecast standard deviation σ_g increases. In this example, $\lambda = 1$ and $g \sim N(0, \sigma_g^2)$. The curve has an asymptote at $\sqrt{2\lambda T} = \sqrt{30}$.

By symmetry, the rate of no-shows generated after T is equal to the rate generated before 0, so that the total rate of no-shows, λ_n, is:

$$\lambda_n = 2\int_T^\infty P_n(t)\lambda dt. \tag{2.10}$$

If the density is symmetric about a mean of zero, then the expression 2.10 reduces to equation 2.6. Therefore, the distribution of no-shows is also Poisson with parameter λ_m. Intuitively, each surprise (no-show) is a no-show (surprise) in another period. As with the surprises, the expected number of no-shows have an upper limit of λT as σ_g increases. This is the expected number of arrivals during the demand period of length T (see Figure 2-4).

Demand Forecast Error (e)

Since arrivals appear according to a Poisson process and arrival time forecast errors are independent, the number of surprises and no-shows are also independent. Therefore, the distribution of the demand forecast error $e = n - m$ may be found by convoluting the distributions of n and m. The mean of e is zero (we expect equal numbers of surprises and no-shows as long as the arrival process is homogeneous), while the variance is the sum of the variances of surprises and no-shows. Under the assumption of independent, normal, zero-mean forecast errors,

$$\sigma_e^2 = \sigma_m^2 + \sigma_n^2$$

$$= 2\lambda_m$$

$$= 4\lambda[T(1 - F_g(T)) + \sigma_g^2(f_g(0) - f_g(T))] \quad (2.11)$$

and as σ_g grows, $\sigma_e \rightarrow \sqrt{2\lambda T}$. When individual arrival-time forecasts have a large variance, the variance of the demand forecast error approaches twice the expected number of arrivals during the demand period. The standard deviation, σ_e, is displayed in figure 2-5 as a function of the forecast standard deviation σ_g.

If we assume that arrival-time forecasts have a standard deviation of 10 minutes, then the standard deviation of the demand forecast during one 15-minute period is 3.8 aircraft (see Figure 2-5). We expect 15 aircraft during the demand period, so the standard deviation is approximately 25% of the mean. This represents a large amount of uncertainty injected into traffic management decisions.

Figure 2-5 also indicates that the standard deviation of the demand forecasts fall rapidly as arrival time forecast standard deviation falls below 10. Therefore, there is much to be gained by reducing take-off time uncertainty below its present level.

2.3 Conclusions and Extensions

The previous demand model assumed a homogeneous poisson arrival rate and independent arrival forecast errors. Additional analyses may relax these assumptions, or a simulation may be used to find the demand forecast error rate for an actual arrival schedule. Also of interest is the pattern of forecast errors over multiple periods. Errors among periods are correlated, since surprises in one period are no-shows in another. Given a finite number of flights over a limited number of demand periods, the distribution of demand forecast errors over time periods might be expressed in terms of combinatorial probability (i.e. an expression similar to the multinomial distribution).

In addition, these calculations may not provide a full description of the effect of forecast errors on strategic traffic management decisions. Flights which are no-shows or surprises usually appear in a period *adjacent* to the expected period of arrival. In other words, normally distributed forecast errors with a small variance produce local fluctuations in actual demand around the projected demand, and these local effects would not have a great impact on the effectiveness of global flow management strategies. Temporary saturation of

airport capacity can be alleviated by path-stretching or speed control of arriving aircraft. It is preferable not to resort to these tactical flow management procedures, but they have less of an impact when compared to the length of delays and number of aircraft involved in a typical GHP. Further research would evaluate the effect of these uncertainties on the performance of Ground Holding Programs in particular and on air traffic management actions in general.

Chapter 3

Description of Data Sources and Preparatory Data Analysis

The aircraft departure models developed in this thesis must be tested against real data, and the collection and consolidation of the necessary data presented many challenges. At this time, the FAA does not maintain a central database with information about all operations at major airports. Nor does the FAA regularly receive either real-time or historical information from the carriers. Therefore, we gathered data from the FAA, the National Weather Service, the Logan Airport tower, and three major carriers in order to construct a data set for fitting and testing the models. We were able to 'match' flights in distinct data sets in order to build more complete records of each aircraft's trajectory. However, complete records for some subsets of flights could not be constructed. This chapter describes the procedures used to assemble the data set and provides summary statistics of the data. Potential limitations and sources of bias in the data are also described. The final section of the chapter presents recent improvements to the FAA's data collection capabilities

In subsequent chapters, predictions of ground transit times (GTTs) will be derived from the observed ground transit times of recently departed aircraft as well as from predictions of environmental factors, such as the weather. Therefore, there are two general categories of data which have been collected:

- *Flight Data* contain information about particular flights, such as the scheduled and actual take-off times.
- *Environmental Data* contain information about factors which affect a number of flights

during a significant time period, such as weather conditions and runway configuration.

Section 3.1 discusses the flight data and the procedures to match flights in multiple databases. Section 3.2 describes the environmental data while Section 3.3 examines potential problems due to sampling bias or 'noise' in the data. Finally, Section 3.4 describes current FAA inititiatives to improve data collection and analysis.

Data from these two categories will be described in the following sections.

3.1 Flight Data

We obtained three sets of flight-by-flight data: a historical database collected by the ETMS itself, the DOT's Air System Quality Performance (ASQP) data, and the Logan Airport gate schedules of three major carriers. Table 3.1 describes the contents of these databases. The second column of the table lists the elements of a flight's record in each database and the final column describes the sample of flights contained within each database and the time period covered by the data. Records from the ETMS data contain information about actual aircraft take-off times, while the ASQP data provide actual times of departure (push-back) from the gate and arrivals to the gate. In addition, the ASQP data contain the scheduled (published) departure times while the ETMS data contain the proposed departure time of each flight according the to the filed flight plan. Connections between arriving and departing aircraft may be inferred from the gate schedule data. Gate schedules were contributed by American Airlines as well as two additional carriers which wished not to be identified. These two will be labeled 'Carrier B' and 'Carrier C.' The third column indicates that only major carriers operating at Logan are covered by both the ETMS and ASQP data, while only August flights by Carrier B are covered by all three databases.

Ground transit times may be broken into gate departure delay and roll-out time (see Figure 1-1 in Chapter 1). The ETMS data contributed GTTs of departures from Logan, while gate departure delays were calculated from the ASQP data. Flights found in both the ETMS and ASQP will be called *matched* flights. For matched flights, roll-out times are calculated by subtracting the gate departure delay from the GTT. The following sections describe these data in more detail. The gate schedule information will be described in more detail in Chapter 4 when it is incorporated into a model linking arrival and departure delays.

Database Name	Contents of Record	Sample
ETMS	scheduled gate departure proposed gate departure DZ message (approx. take-off) cancellation indicator flight number carrier aircraft size code	Logan Airport only; All flights; March 5 - 15, 1991; August 16 - 31, 1991.
ASQP	scheduled gate push-back scheduled gate arrival actual gate push-back actual gate arrival cancellation indicator flight number carrier	All major airports; Major carriers only; Jan. 1991 - Dec. 1992.
Gate Schedule	connection flight numbers	Logan Airport only; AA, January - December, 1992; Carrier B, June - August, 1991; Carrier C, December 1992.

Table 3.1: Airport Flight Data

	March	August	Total
number of flights	4245	7343	11588
median	17	21	19
mean	22.4	26.2	24.8
standard deviation	17.7	19.0	18.6

Table 3.2: Summary Statistics of Ground Transit Time for all Recorded Logan Departures During March and August, 1991

3.1.1 Ground Transit Times from the ETMS Data

The ETMS historical database contains records of departures from Logan Airport for 11 days in March and 16 days in August, 1991. Once duplicates and records with missing fields are removed, 4245 complete records remain for March and 7343 records remain for August. Note that the August sample contains proportionally more flights than March. After adjusting for the number of days in each sample, the August data has, on average, 16% more flights on each weekday and 28% more flights on each weekend day. A likely explanation for this increase is the extra traffic between Logan and summer destinations such as Cape Cod.

Each flight in the ETMS database contains the proposed departure time listed in the flight's filed flight plan and the time of the aircraft's DZ message. We will use the DZ message as an estimate of the aircraft's actual take-off time. The estimated GTT for each flight in the ETMS data is the difference between actual take-off and scheduled departure:

$$\text{estimated GTT} = \text{DZ message time} - \text{proposed departure time}$$

The first graph in Figure 3-1 displays GTTs on March 11, a Tuesday with snow flurries in the morning and fair weather in the afternoon. Each '*' represents the GTT of one flight. Note the extremely high variability of GTT, even during the afternoon. Table 3.2 contains summary statistics for the sample of ground transit times. While the overall mean GTT was approximately 25 minutes, 6% of the GTT were above one hour while 1% were below zero. Aircraft with negative ground transit times would have left the gate before the proposed departure time. This is not unusual for General Aviation and freight carriers.

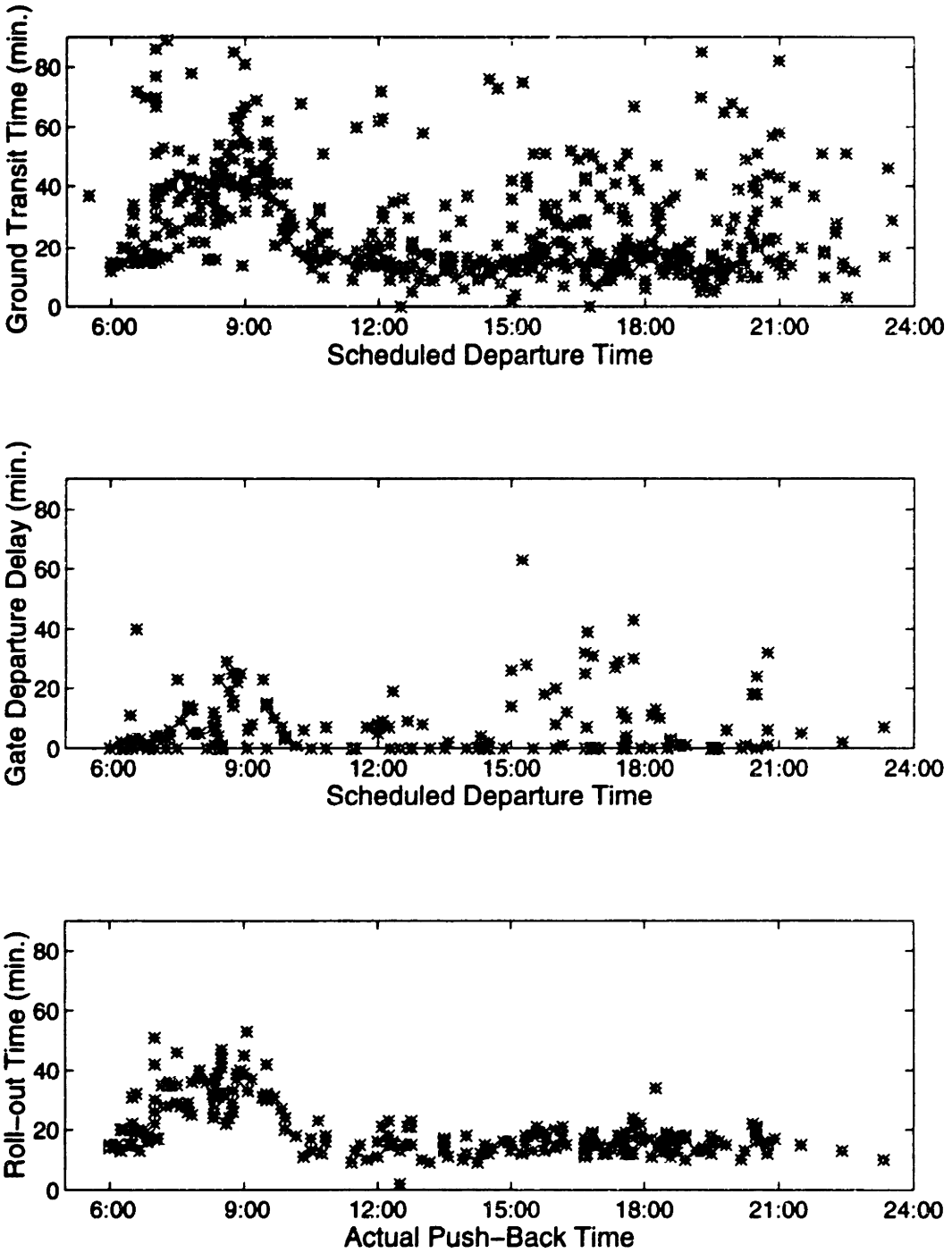


Figure 3-1: Ground transit times, gate departure delays, and roll-out times of individual flights at Logan Airport on Monday, March 11, 1991. The first graph displays all flights in the ETMS data, while the second and third graphs only show 'matched' flights found in both the ASPQ and ETMS data.

3.1.2 Gate Departure Delay and Roll-out Time from the Matched Data

In order to calculate roll-out times for individual flights, departures in the ETMS and ASQP data were matched. For each matched flight, GTT and gate departure delay were calculated from the ETMS and ASQP data, and their difference is the roll-out time. The following criteria were used to match a record from the ETMS database with a record from the ASQP database:

- Records in both databases must be complete. This eliminated damaged records and canceled flights.
- The records must indicate the same origin, destination, carrier, and flight number.
- The scheduled departure time in the ASQP database and the proposed departure time in the ETMS database must be within six hours of each other. Therefore, the published departure time and the departure time in the filed flight plan must be within six hours of each other in order for the two to be considered the 'same' flight. This eliminates the possibility that flights on two different days may be matched.

Flights that satisfy all of these criteria will be referred to as *matched flights*, and the resulting data as *matched data*. Overall, 44% of the flights in the ETMS database were matched (see Table 3.3) while 75% of the flights in the ASQP database were matched.

The low percentage of ETMS matches can be attributed, in part, to Logan's traffic mix. The ETMS contains a cross-section of all flights, while the ASQP database contains flights operated by the major carriers. According to FAA records, in 1991 approximately 234,000 out of the 441,000 operations at Logan were performed by the major carriers [17]. This is 53% of all operations; the remaining operations were performed largely by air taxis and commuter airlines (40%) and general aviation (7%). Since the ASQP data includes only the ten largest domestic air carriers, the 44% match rate is not surprising.

Table 3.3 also shows that 25% of the ASQP flights were not found in the ETMS data. Most of these unmatched flights departed during 'gaps' in the ETMS data. During these gaps, no flights were recorded by the ETMS, even though the ASQP data indicate that the airport was operating normally. ETMS personnel believe that these gaps were caused by computer 'down-time' for maintenance or re-programming and that these gaps are not related to the status of the air traffic system.

	number of flights		
	March	August	Total
ETMS Historical Database	4245	7343	11588
ASQP Database	2739	4020	6759
Matched Data	2043	3002	5045
% ETMS Matched	48%	41%	44%
% ASQP Matched	75%	75%	75%

Table 3.3: Number of flights recorded and matched during the sample period, March 5-15 and August 16-31, 1991.

		March	August	Total
	number of flights	2043	3002	5045
gate departure delay	median	0	0	0
	mean	4.2	6.2	5.4
	standard deviation	12.5	15.0	14.1
Roll-out time	median	16	17	17
	mean	17.2	19.1	18.3
	standard deviation	8.1	8.5	8.4
ground transit time	median	17	20	19
	mean	21.4	25.3	23.7
	standard deviation	15.1	17.8	16.8

Table 3.4: Summary statistics of gate departure delay, roll-out time, and ground transit time for matched data (flights in both ETMS and ASQP databases)

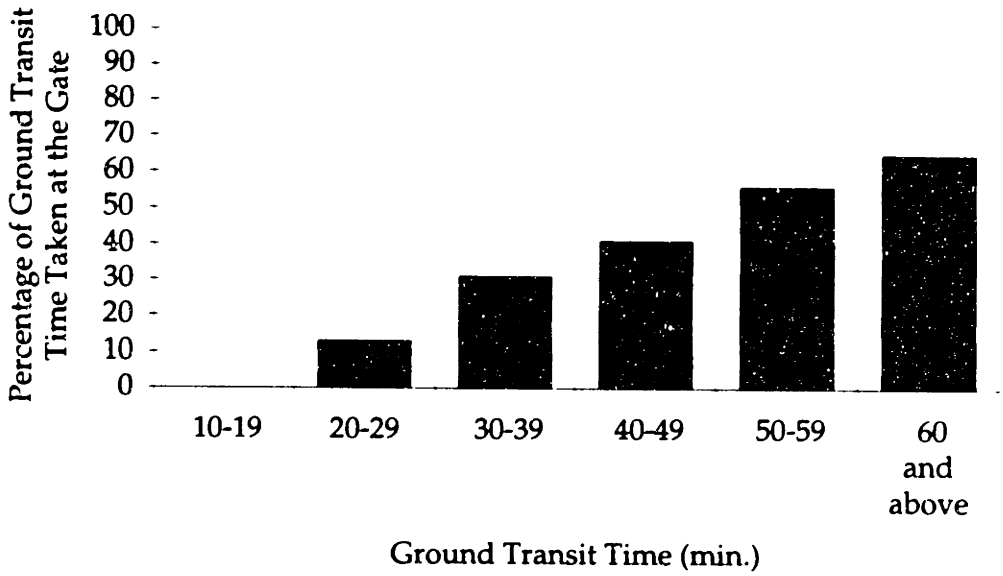


Figure 3-2: Percentage of ground transit time taken at the gate for increasing levels of ground transit time

Table 3.4 contains summary statistics for the gate departure delays and roll-out times for matched flights, and Figure 3-1 shows gate departure delays and roll-out times on March 11, 1991. The table indicates that the mean roll-out time of 18 minutes is significantly higher than the mean gate departure delay of 5.4 minutes, but that the variance in gate departure delay is much higher than the variance in roll-out time. Since gate departure delays have a long 'right tail', the mean gate departure delay of 5.4 minutes is larger than the median gate departure delay of 0 minutes. This is also seen in Figures 3-1; roll-out times have a low variance during the afternoon, and the high variance in GTT is contributed primarily by the gate departure delays.

On average, gate departure delay represents under one-quarter of total GTT. However, the impact of gate departure delay varies with the size of the total delay. As GTT grows, the percentage of delay taken at the gate grows as well. Figure 3-2 displays the percentage of delay taken at the gate as the total ground transit time varies. For GTTs between 10 and 20 minutes, gate departure delays were negligible. For GTTs above one hour, gate delay contributed 65% of the total. Therefore, attention should be given to predicting both roll-out times and gate departure delays even though gate departure delays are, on average, a much smaller contributor to overall GTT.

3.2 Environmental Data

Since we expect ground transit times to be influenced by both weather and runway configuration, the matched data described in the previous section were combined with data from the National Weather Service (NWS) and Logan Airport's Preferential Runway Advisory System (PRAS). The NWS data contain hourly observations of ceiling height, visibility, wind speed and wind direction at Logan airport. The ceiling height and visibility were used to classify each hour into one of four standard weather categories: Visual Flight Rules (VFR), Marginal VFR (MVFR), Instrument Flight Rules (IFR), and Low IFR (LIFR). The categories are determined by comparing ceiling and visibility to the minimum values shown in Figure 3-3. Each departure in the matched data was assigned the weather category seen at the time of its push-back from the gate. The bottom row of Table 3.5 displays the percentage of all matched flights which occurred under each weather condition.

The PRAS data describes Logan's runway configuration throughout the day. Each departure in the matched data was assigned the active runway configuration at the time of its push-back from the gate. Table 3.5 displays the percentage of the 5045 matched flights which fell under each weather category and runway configuration. A majority of flights, 3274 out of 5054, or 65%, operated under one of three high-capacity runway configurations (4LR/9-4R, 27-22L/22LR and 33LR/27-33L) and under VFR. Despite the large number of flights in the matched sample, there are few flights in the sample which operated under the less common runway configurations and the worst weather. Only 3% of the flights (132 flights) operated under LIFR. The low number of LIFR flights in the matched data is due to both the infrequency of severe weather conditions and the cancelation of flights during these conditions.

3.3 Sources of Error and Bias

The data described above will be used to test the forecasting procedures, and errors in the data may affect the results of the empirical tests. There are two potential problems in the data: measurement error and sampling bias. The former problem is caused by the use of the DZ message as a proxy for take-off time. The latter problem may stem from the small sample period as well as the selection bias introduced when only matched data are used for testing. In this section we discuss the causes and potential effects of these errors. We

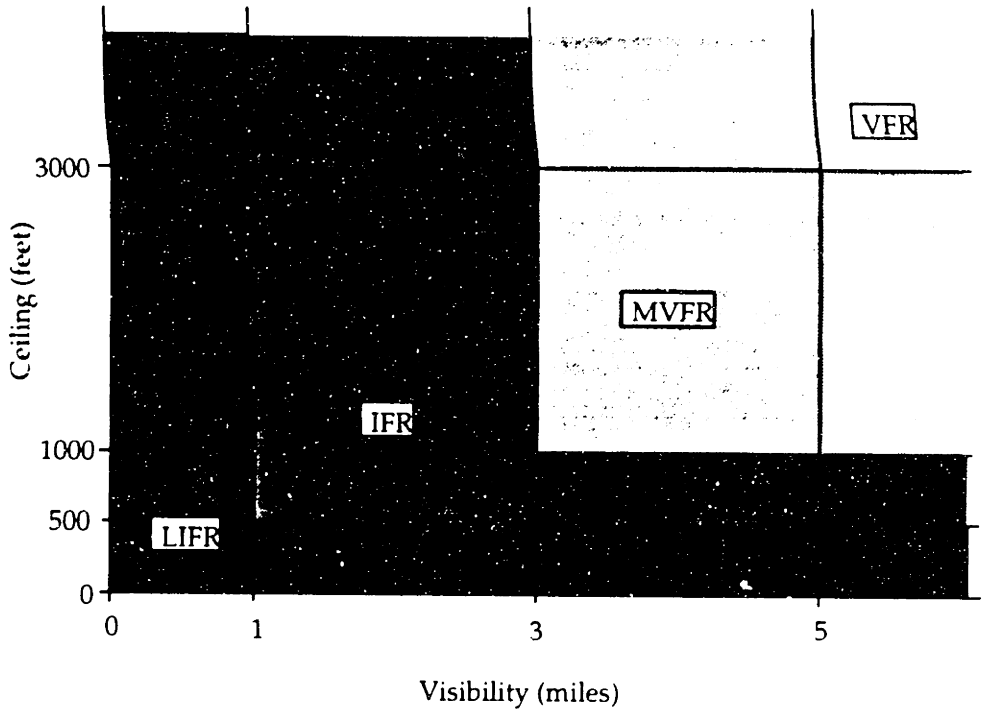


Figure 3-3: Standard weather categories

	VFR	MVFR	IFR	LIFR	total
4LR/9-4R	23	2	3	2	30
22LR/22LR	3	2	1	0	6
27-22L/22LR	19	6	2	0	27
33LR/27-33L	23	3	2	0	29
33LR/33LR	5	1	1	0	6
Other-Runway	0	0	1	0	1
total	73	14	10	3	100

Table 3.5: Percentage of 5045 matched flights operated under Logan runway configurations and standard weather categories (columns and rows do not sum to margins because of rounding)

determine that while these problems exist, they should not affect fundamental results of the thesis.

Use of the DZ message as an estimate of take-off time may introduce errors into our estimates of ground transit times. According to the *ETMS Functional Description*, a DZ message "signifies the activation of a proposed flight" in the National Airspace System [39]. Gary Hofnagle, an engineer in the Boston Tower, adds that the DZ message is generated when a departing aircraft's beacon first responds to an interrogation from the tower's Secondary Surveillance Radar (SSR) [21]. When the message is generated, a filed flight plan is activated and a flight strip is produced. However, portions of the airfield are masked from the SSR in order to minimize the radar noise from objects on the ground, and the DZ message may be delayed until after the aircraft has climbed above the masking.

While Mr. Hofnagle estimated that the lag between take-off and DZ message generation should not be more than ten seconds, comparisons between DZ messages and actual take-off times have revealed that DZ messages at some airports may be generated almost 5 minutes after actual take-off, on average [36]. Unfortunately, no detailed information on the difference between DZ message and actual take-off time is available for Logan Airport, so that it is not possible to adjust for possible errors in the data. However, it seems reasonable to treat take-off time error as random noise of the sort found in almost any electronic monitoring system. We will not explicitly correct for this error, but it should be remembered that variations in take-off times may be due, in part, to deviations of the DZ messages from the true take-off times.

Sampling bias may also be of some concern, for it is not likely that a sample of flights from 27 days at Logan is representative of flights at other airports or even at Logan at other times of the year. The sample itself contains some unusual conditions: August 18 - 20 saw the passage of Hurricane Bob up the East Coast. While we cannot be sure that the results derived in this thesis would apply throughout the air traffic management system, the data do provide a thorough test of the forecasting procedures. Both VFR and IFR conditions are well-represented in the sample, and empirical tests with these data should determine how the forecasting procedures are affected by good, bad, and extreme weather at Logan airport. The data will also provide a good test of the relative merits of the forecasting procedures.

Empirical tests of roll-out time forecasts will be performed on flights in the matched

data only, since roll-out times are only known for these flights. The results of these tests may not be representative of the results that would be obtained if all flights were used since the distribution of roll-out times for matched flights may be significantly different from the distribution for unmatched flights. We cannot compare their roll-out times directly, but we can compare their ground transit times. Table 3.6 contains GTT summary statistics for matched and unmatched flights and Figure 3-4 displays the empirical cumulative distributions of the GTTs of each sample. The table and the figure suggest that while both samples have similar means, the GTTs of the unmatched flights have a higher variance than those of the matched flights. Nonparametric statistical tests confirm these observations. The Wilcoxon rank-sum test produces a statistic that indicates a difference in location between the distributions which is almost, but not quite, significant at the 0.05 level (z-stat $= -1.6$, p-value $= 0.06$). However, a two-sample Kolmogorov-Smirnov test unequivocally rejects the hypothesis that the two empirical distributions shown if Figure 3-4 were produced by populations with the same distribution of GTT (z-stat $= 5.07$, p-value < 0.0001). A large difference in dispersion, rather than location, distinguishes the two samples.

Without more information about the unmatched flights, we cannot be sure whether the differences in Ground Transit time are due to differences in gate delay, roll-out time, or both. We do know that only the major carriers are included in the matched data, while the unmatched data are dominated by commuter and general aviation aircraft. It is likely that aircraft types, push-back procedures, and scheduling practices vary between these subsamples, and therefore there may be differences in both gate departure delays and roll-out times. However, we feel that the bias introduced by the use of matched data is not of great concern. First, the differences between matched and unmatched flights, while statistically significant, are not enormous (i.e. a 15% difference in standard deviation). In addition, the FAA may be primarily interested in take-off time predictions of flights by the major carriers (the only flights found in the matched data), for these flights are most often subject to major air traffic management procedures, such as ground hold programs.

3.4 Improving FAA Data Collection: Current Initiatives

This analysis has highlighted the scarcity of clean and complete data on airport operations. Without supporting data, empirical analysis and statistical model-building are impossible.

	Matched	Unmatched
Number Flights	5045	6543
Mean (min.)	23.7	25.6
Median(min.)	19	20
Standard Deviation (min.)	16.8	19.8

Table 3.6: Summary statistics for ground transit times of matched and unmatched flights

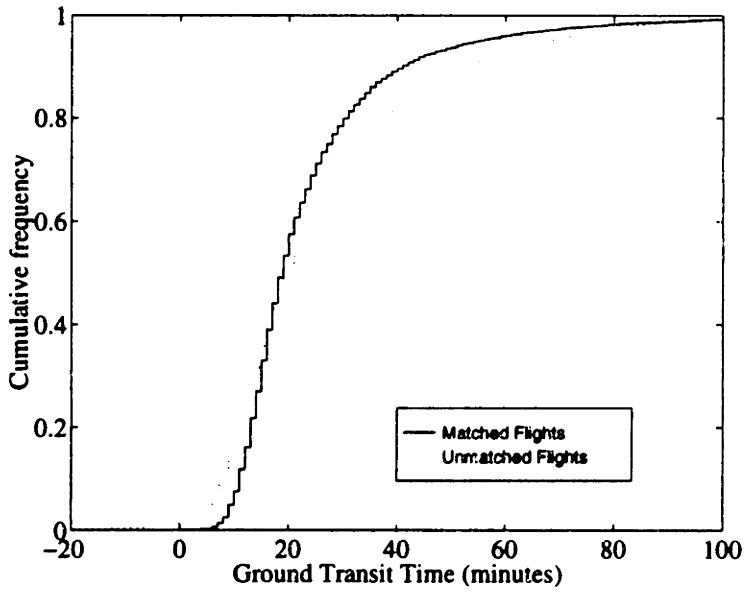


Figure 3-4: Cumulative distributions of ground transit times of matched and unmatched data.

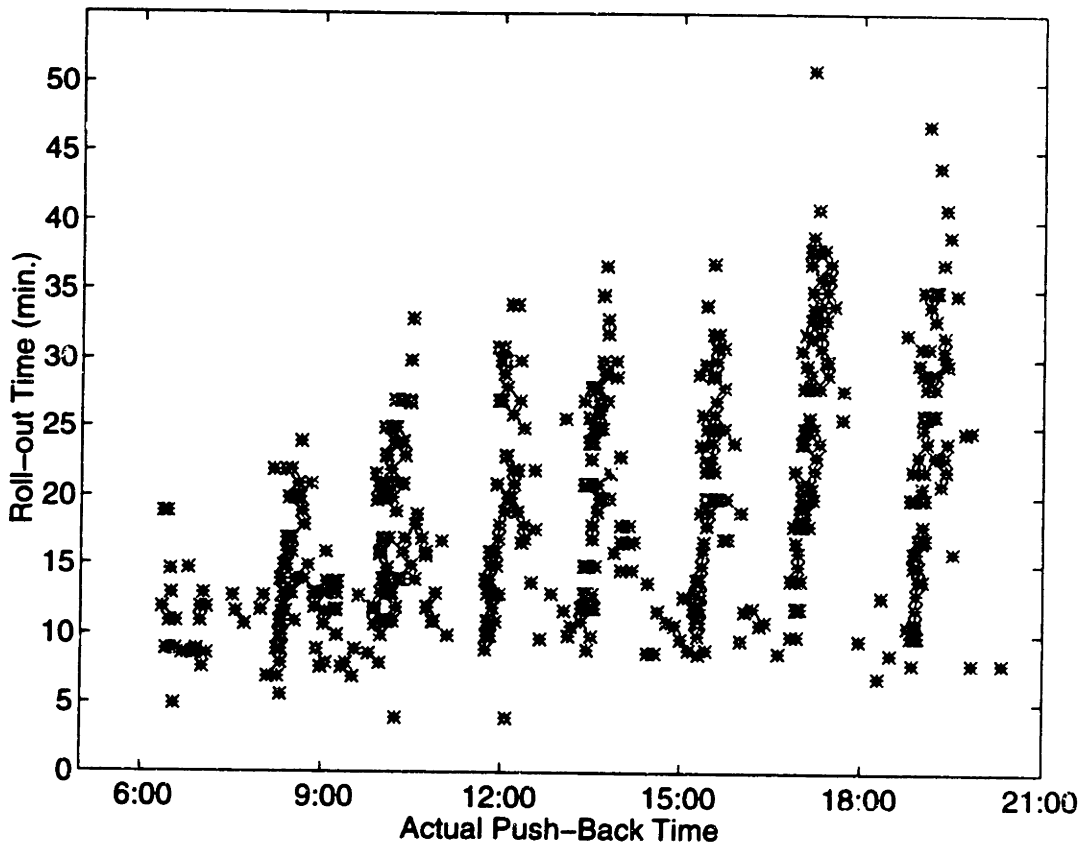


Figure 3-5: Roll-out times of flights from Atlanta Hartsfield International Airport on August 1, 1993.

We have assembled a data set which will allow us to asses the performance of our forecasting models under a wide variety of weather conditions and runway configurations at Logan Airport. The FAA has also recognized the need for such a data set, and two recent projects have improved the FAA's resources. While both data sets are similar to our 'matched' data, neither were available in time for an analysis to be included in this thesis.

The first new data set is an upgraded version of the ASQP which contains records for flights by the ten major carriers from January, 1995. Each record in the ASQP now includes actual push-back and gate arrival times [14]. Therefore, each record may now be used to generate gate departure delays and roll-out times. The second database is the Consolidated Operations and Delay Analysis System (CODAS). The CODAS data set is similar to our own data set and was also constructed by matching historical ETMS and ASQP records [1]. The CODAS data were gathered for flights which operated during 1992 - 1993 at almost 100 of the largest domestic airports. As with the new ASQP and our own

database, the CODAS data contain gate departure delays and roll-out times. Figure 3-5 shows CODAS roll-out times of flights from Atlanta Hartsfield International. A comparison with Figure 3-1 demonstrates that the pattern of delays is strikingly different from the pattern at Logan and is characterized by increasing roll-out delays within, and between, large banks of departures. Clearly, there are rich opportunities for analysis and modeling of departure delays at airports besides Logan.

While the new ASQP and CODAS data sets are improvements over previous data sources, neither offer a full description of airport operations. As with the matched data, these data sets only contain flights operated by the ten largest carriers. One advantage of our 'home-grown' data set is that the unmatched flights are also available for analysis. This is especially important at Logan, where many flights do not show up in either the ASQP or CODAS data. For example, in Chapter 5, departure queue sizes will be derived by estimating the push-back times of all flights and then counting the total number of aircraft on the airfield. Using CODAS or ASQP alone, approximately half the aircraft would not be counted in the total.

In the long run, the FAA should expand its databases to include real-time information collected from major air traffic facilities and the air carriers. However, there are a multitude of possible sources of data, and the FAA will have to establish priorities. Experiments such as the ones conducted in this thesis will point out which data are most likely to produce significant benefits. The importance of one data source, the carriers' real-time delay and connection information, will be highlighted in the next chapter.

Chapter 4

Forecasting Gate Departure Delay

In the previous chapter we saw that the average gate departure delay (GDD) in the matched data is just above five minutes, while the average ground transit time (GTT) is 24 minutes. However, GDDs have a high variance and represent a large percentage of the total GTT when flights experience long delays. If a forecasting model can be devised which accurately predicts GDD, we will have made substantial gains in reducing take-off uncertainty.

Since departures may be delayed at the gate by a large number of factors, accurately predicting GDD is a difficult problem for both the carriers and the FAA. The delay reporting system managed by American Airlines indicates the complexity of the problem. Each delay is classified with one of over one hundred possible delay codes. The list includes codes for weather, mechanical, and crew problems, as well as codes for the late arrival of a dignitary and the removal of flying insects from the cockpit [22]. Many of these delays occur irregularly and are nearly impossible to predict, and it is inevitable that there will be some irreducible uncertainty in GDD forecasts. This chapter explores whether improved forecasting procedures may significantly reduce the current level of uncertainty.

Important sources of information on GDDs are the carriers themselves, for they continuously monitor the movements of aircraft and crews. Section 4.1 describes previous research on the accuracy of the carriers' own delay forecasting systems. Section 4.2 examines our database of matched flights to determine whether knowledge of factors such as weather or runway configuration can significantly improve GDD forecasts. We also determine whether GDDs and roll-out times are correlated, for a high correlation would suggest that a model which integrates GDD and roll-out time predictions would improve forecast accuracy. Sec-

tion 4.3 describes a simple model for GDD that describes the 'turns' made by individual aircraft. This model is successful in predicting many of the largest gate departure delays, given accurate predictions of arrival delays. The chapter concludes with suggestions for more detailed models and improved data-sharing between the FAA and the carriers.

4.1 Gate Departure Delay Forecasts by the Carriers

Recent projects by the FAA and the MITRE corporation analyzed the real-time schedules of the major carriers in order to determine the accuracy of the carriers' own departure forecasts [31], [10]. The real-time schedules contain the carriers' forecasts of aircraft push-back times. For some delayed flights in the sample, these forecasts were altered as the time of push-back approached. The forecasts for these flights were *updated* when the carrier anticipated a delay. For most flights, however, the carriers retained the original scheduled push-back times until the actual push-back occurred; the original forecast was not updated.

Under ideal circumstances, these data may have been used to determine whether the carriers' updated forecasts were more accurate than the forecast algorithms developed for the ETMS. In fact, the studies found that the updated schedules were only marginally more accurate than the original, published schedule. However, it is likely that some of the "real-time schedules" submitted by the carriers were the departure time forecasts that had been sent to gate agents and the passengers themselves [11]. These predictions may have been influenced by the carriers' customer service objectives and may not have been the carriers' most accurate forecasts. For example, forecasts released to passengers may have underestimated the longest delays so that customers would be back at the gate and prepared to board if the aircraft were ready more quickly than was anticipated. It is likely that the carriers do produce more accurate, unbiased forecasts for operational decisions, but these have not been made available to the FAA.

In addition to the real-time schedules, MITRE also received the delay codes for those flights with GDDs larger than 5 minutes. The delay codes were grouped into seven categories: delays due to weather, service problems, passenger delays, mechanical difficulties, crew duties performed by pilots and flight attendants, air traffic control, and other miscellaneous delays. One category of delay that was *not* reported was delays due to late arrivals. Therefore, 'upstream' delays are not included in the breakdown even though this is likely

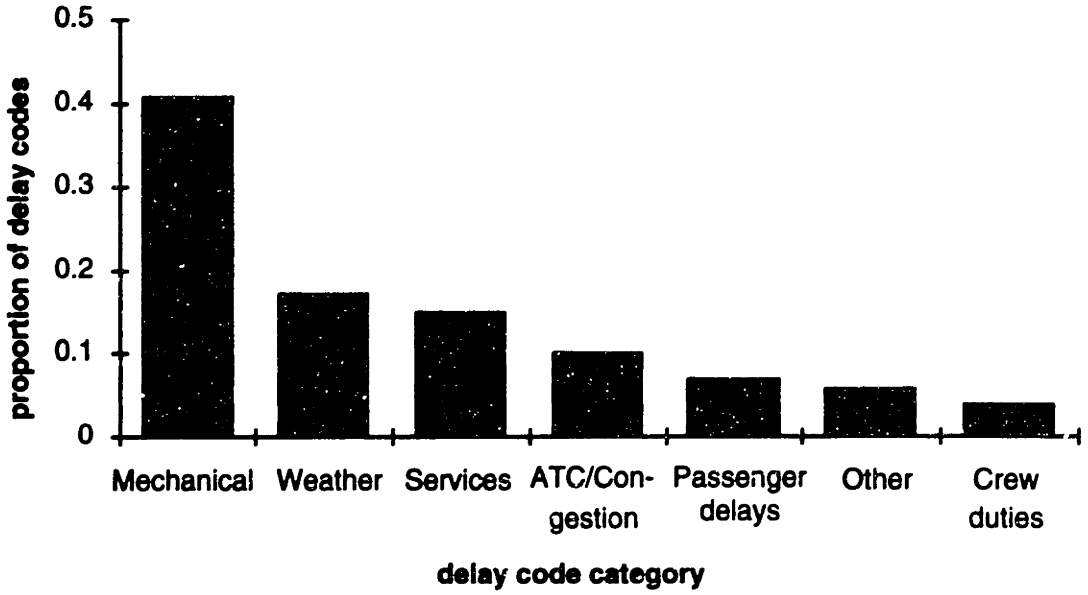


Figure 4-1: Proportion of delay codes for flights with updated forecasts

to be a significant, and relatively predictable, cause of delays.

The delay codes were further broken down into codes for delayed flights that did receive forecast updates, and codes for delayed flights with departure time forecasts that were not updated. The proportions of delay codes in each group are shown in Figures 4-1 and 4-2. Large proportions of the reported delays were due to mechanical problems, passenger delays, and crew duties. Flights with push-back times that were updated in the real-time schedule, however, had proportionally more mechanical delays, while those that were not updated were most often delayed by passengers or crew. This disparity may be caused by the relative length of mechanical delays (for example, the carrier would want to inform its passengers about an hour delay), but the carriers' delay reporting policies are not known. In general, the most common delays in Figures 4-1 and 4-2 are difficult for a centralized forecasting system to predict.

Another interesting element of the FAA/MITRE study is the timeliness of the forecast updates. Over 70% of the forecast updates arrived within one hour of the scheduled departure time of the flight; only 5% of the updates arrived over three hours in advance. If the carriers were supplying their most up-to-date data, then the value of carrier delay information for long-term forecasts may be limited. Most of these carrier forecasts only diverge from the scheduled departure time when the scheduled departure time is close. It is quite possible, however, that the studies did not include the most timely carrier forecasts.

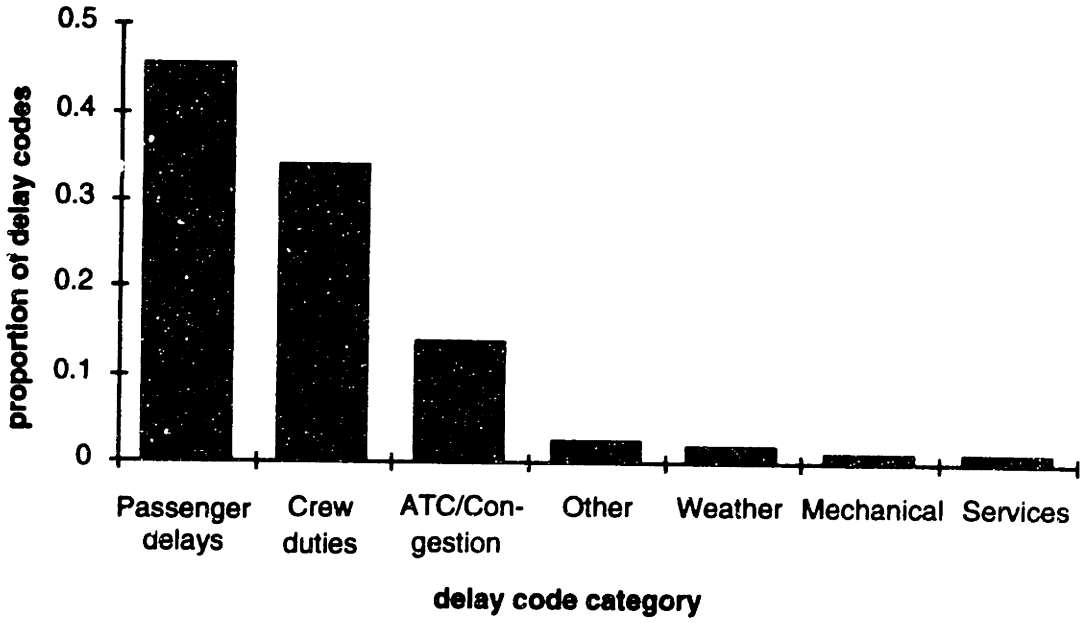


Figure 4-2: Proportion of delay codes for flights with no updated forecasts

4.2 Explanatory Variables for Gate Departure Delay

One would expect that weather and runway configuration would be useful for predicting GDD. One might also expect GDD to be correlated with roll-out time, for conditions which lead to push-back delays may also lead to roll-out delays. In this section we investigate whether GDD is significantly correlated with these factors and, if it is, whether models built from these factors are successful in reducing the uncertainty associated with predicting GDD.

4.2.1 Environmental Factors

Airline passengers often find that bad weather means long delays. The gate delay data confirm this observation. The mean GDDs under three weather categories are displayed along the bottom row of table 4.1 (because of the relatively small number of LIFR flights, this category has been combined with IFR). Particular runway configurations, such as 4LR/9-R and 27-22L/22LR are associated with the longest GDDs in bad weather. This is counterintuitive, for these two runway configurations have the *highest capacity* of those runway configurations listed and are equipped with the most sophisticated aircraft guidance systems. One explanation is that the guidance systems of these configurations are needed when the weather is at its worst, and both arrival and departure delays are particularly

	weather category			
Runway Config.	VFR	MVFR	IFR/LIFR	total
4LR/9-4R	3.7	16.1	21.4	7.7
22LR/22LR	4.7	5.1	2.9	4.4
27-22L/22LR	5.5	5.2	6.3	5.5
33LR/27-33L	3.1	4.5	9.3	3.7
33LR/33LR	2.1	8.4	2.5	3.0
Other	1.6	5.6	5.5	4.2
total	3.9	6.8	12.7	5.4

Table 4.1: Mean GDD by Runway Configuration and Weather Category (min.)

long at these times.

These differences in means are not statistical aberrations caused by the overall variance in GDD. The nonparametric Kruskall-Wallace test indicates that the locations of the GDD distributions vary significantly as weather conditions and runway configurations change. The Kruskall-Wallace test assigns ranks to the entire sample of GDDs and then averages the ranks of GDDs which occurred under each factor. For example, the ranks of delays under VFR, MVFR and IFR/LIFR are each summed and the totals divided by the number of observations in each group. If the rank averages diverge, then this is evidence that the delays under each factor do not share a common location. This test rejected the null hypothesis that the distribution of GDDs under varying conditions had the same locations (p-values < 0.0001 for both weather categories and runway configuration).

While the effects of weather and runway configuration are statistically significant, the tests described above do not demonstrate how valuable these environmental factors will be for predicting GDD. If observed GDDs can be closely approximated by some function of weather state and runway configuration, then this function may be used to predict GDDs, given predictions of weather and runway. Consider the following linear model for GDD:

$$g_{ijk} = b + w_j + c_k + u_{jk} + e_{ijk} \tag{4.1}$$

where

g_{ijk} = the GDD of the ith flight operating under weather category j and runway configuration k

b = mean (or 'base') GDD applied to all departures

Group	Sums of Squares
Weather	$s_w^2 = 43,475$
Runway Config.	$s_c^2 = 13,233$
Interactions	$s_u^2 = 39,873$
Sum of Squared Errors	$s_e^2 = 902,602$

Table 4.2: Results of Analysis of Variance for GDD Classified by Weather and Runway Configuration

$$
\begin{aligned}
w_j &= \text{additional GDD due to weather category } j \\
c_k &= \text{additional GDD due to runway configuration } k \\
u_{jk} &= \text{additional GDD due to interaction between weather } j \text{ and runway } k \\
e_{ijk} &= \text{residual GDD not explained by model factors}
\end{aligned}
$$

Let $\bar{b}, \bar{w}_j, \bar{c}_k$, and $\bar{u}_{jk}$ be estimates of b, w_j, c_k, and u_{jk} obtained from the data. The observed model errors $\hat{e}_{ijk}$ are:

$$\hat{e}_{ijk} = g_{ijk} - (\bar{b} + \bar{w}_j + \bar{c}_k + \bar{u}_{jk}) \tag{4.2}$$

The estimates which minimize the sum of squared errors $\hat{e}_{ijk}^2$ can be derived from the means of the GDDs, grouped by the appropriate factors. For example, the optimal estimate for b under the squared error criteria is the overall mean GDD. The efficacy of this model is summarized in table 4.2. The table contains the sum of squared errors, s_e^2, as well as the portion of the squared difference from the mean 'explained' by the terms in the model. For example, let n_j be the number of observations in weather group j, and recall that there are three weather groups. Therefore,

$$s_w^2 = \sum_{j=1}^{3} n_j \bar{w}_j^2 \tag{4.3}$$

That is, the term s_w^2 is the sum over all observations of the squared differences between the weather category group means $\bar{w}_j$ and the overall mean $\bar{b}$. The sum of squares s_c^2 and s_u^2 are calculated in a similar manner. The sum s_e^2 is the sum of $\hat{e}_{ijk}^2$ over all observations. The table indicates that weather categories, runway configurations, and interactions between weather and runway all contribute to the explanatory power of the model.

Sample	ρ_s (Spearman)	z-value	p-value
All Flights	0.03	2.04	0.04
VFR	0.03	1.60	0.10
MVFR	-0.01	-0.01	0.96
IFR/LIFR	-0.14	-3.46	.0005

Table 4.3: Results of tests for independence between gate departure delay and roll-out time

The standard F-tests of significance for these results are not included in the table, for the data do not satisfy the assumptions necessary for these tests to be statistically valid. Deviations from the model are not normally distributed (they have an extremely long right tail) and the error variance varies between groups. A transformation of the data may alleviate these problems. However, the nonparametric Kruskall-Wallace tests described above indicate that weather and runway are statistically significant factors. A more important lesson from the table is the relative explanatory power of the factors. When compared to the model errors, the weather, runway, and interaction effects are small. One statistic which demonstrates the relative importance of these factors is the R^2 statistic:

$$R^2 = \frac{s_w^2 + s_c^2 + s_u^2}{s_w^2 + s_c^2 + s_u^2 + s_e^2} \approx 0.10 \tag{4.4}$$

This low R^2 value indicates that there is much residual uncertainty about GDDs. There are large GDDs which are not adequately explained by equation 4.1.

4.2.2 Correlation Between Gate Departure Delay and Roll-out Time

If GDD and roll-out time are significantly correlated, then models for the two should be integrated. Given a strong correlation, then an effective model for one should lead to insights on the behavior of the other. However, we find that while gate delay and roll-out time are correlated, the correlation is weak. Let g_i and r_i be the GDD and roll-out time, respectively, of flight i, let $\bar{g}$ and $\bar{r}$ be the overall means, and let N be the total number of flights in the sample (for our matched sample, $N = 5045$). The traditional measure of correlation is the sample correlation coefficient:

$$\rho = \frac{\sum_{i=1}^{N} g_i r_i - \bar{g}\bar{r}}{\sqrt{\sum_{i=1}^{N}(g_i - \bar{g})^2 \sum_{i=1}^{N}(r_i - \bar{r})^2}} = 0.06 \tag{4.5}$$

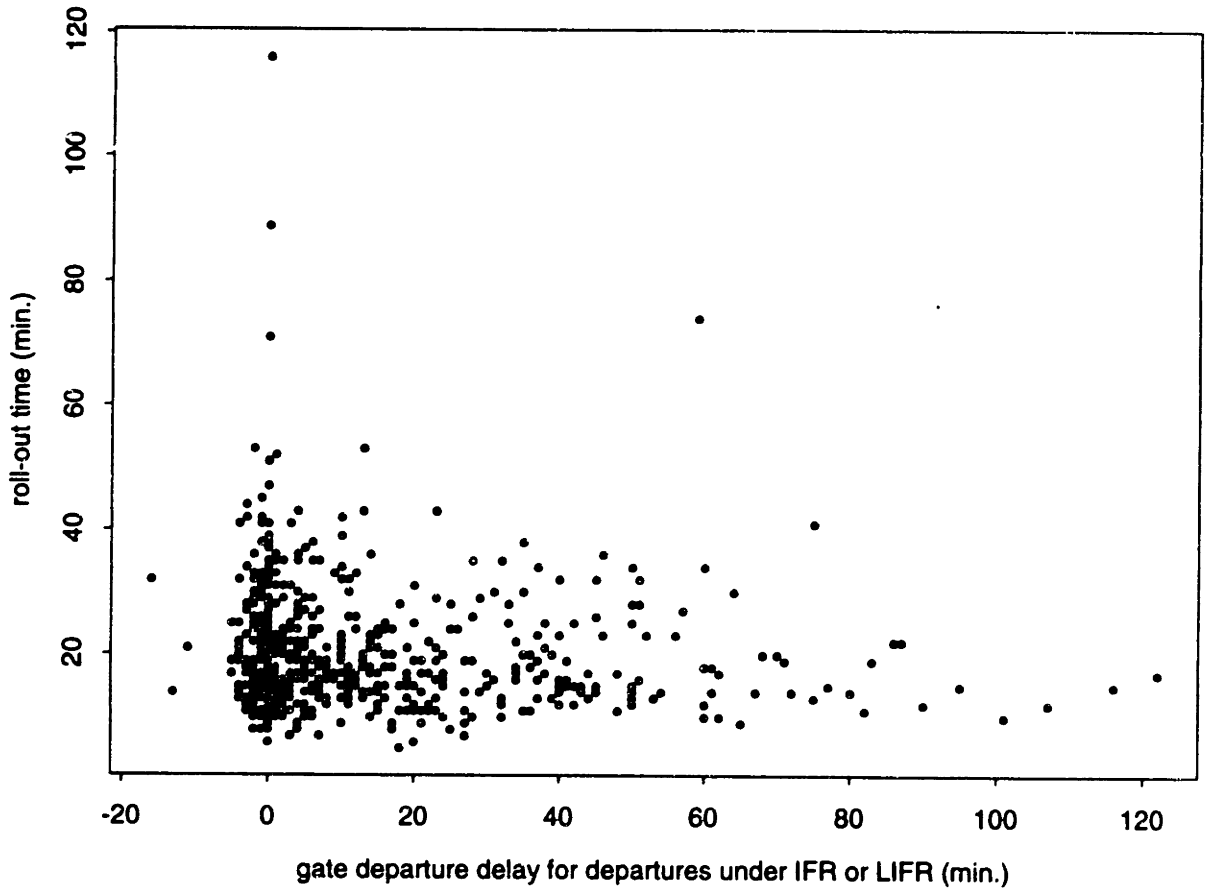


Figure 4-3: Departure gate delays and roll-out times of flights under IFR/LIFR during March and August, 1991.

While this correlation of 0.06 is low, it may be statistically significant. A nonparametric test based on the Spearman rank correlation coefficient, ρ_s, tests whether the correlation between GDD and roll-out time is significant. The coefficient ρ_s is calculated from the difference in ranks of the GDD and roll-out time of each flight. Let G_i and R_i be the ranks of the GDD and roll-out time, respectively, of the ith flight. The Spearman rank correlation coefficient is:

$$\rho_s = 1 - \frac{6[\sum_{i=1}^{N}(R_i - G_i)^2]}{N^3 - N} \tag{4.6}$$

If the ranks R_i and G_i are close for all flights, then ρ_s will be close to one. The maximum possible value of the second term is 2, which is achieved when the ranks of the roll-out times and GDDs are reversed, so that $\rho_s = -1$. Table 4.3 displays the overall ρ_s, as well as ρ_s under particular weather categories. The statistic labeled 'z-value' is derived from ρ_s and is approximately normally distributed. The column labeled 'p-value' displays the significance of the two-sided test of the null hypothesis that $\rho_s = 0$. Overall, there is a significant positive correlation.

A positive correlation between roll-out time and GDD is not surprising, for as airport capacity decreases, both roll-out times and GDD would be expected to rise. The negative correlation under weather categories IFR and LIFR is unexpected, but a close look at the data suggests at least one explanation. The correlation is lowest ('most negative') under particularly severe weather conditions, such as during snow squalls in March and a hurricane in August. Figure 4-3 shows the GDDs and roll-out times of flights during the IFR and LIFR periods in the sample. The negative correlation is produced by the many flights with small GDDs and long roll-out times, as well as the few flights with long GDDs and quick roll-outs.

What caused these estremes? Many of the flights with the largest roll-out times were headed for the same destination during the same bad weather period. For example, there were five flights with roll-out times longer than one hour between 6:00 pm and 8:30 pm on August 18, 1991, and four of these five flights had roll-outs over eighty minutes. Two of these flights were to Newark and three to LaGuardia. In the poor weather, the aircraft may have pushed back on-time to make room at the gate for arriving aircraft. However, it is possible that the subsequent long roll-out times were not spent rolling out, but were spent

on the tarmac waiting for clearance through to the destination airport. Meanwhile, the poor weather may have increased cancellations and arrival delays and slowed the departure rate of flights. Those aircraft that did push-back would find a relatively uncongested runway system and would experience relatively short roll-out times. As a result, some aircraft would have long gate departure delays and short roll-out times and some would push-back on-time but then wait for clearance for departure.

However, without more information, we can only speculate about the causes of the longest delays. Because of the small size of the correlation between GDD and roll-out time, it is reasonable to treat GDD and roll-out time as statistically independent. Therefore, models of gate delay and roll-out times will be developed separately. Forecasts of total GTT may be generated by finding the sum of GDD and roll-out time forecasts.

4.3 Aircraft Turn Models for Gate Departure Delay

In this section we find that a major cause of GDD is the simple fact that an aircraft cannot take off before it lands. An aircraft that arrives far behind schedule will depart late if it does not make its turn before the scheduled departure time. Delays at one airport often stem from delays at an upstream airport.

Define a *turn* to be the operation performed by an aircraft between its arrival to and departure from an airport gate (note that we use the term 'turn' whether or not the flight changes its flight number between arrival and departure). The *scheduled turn time* of a flight is the time between its scheduled arrival at the gate and its scheduled departure from the gate (see figure 4-4). The *available turn time* is the time between actual arrival and scheduled departure. Note from figure 4-4 that increasing arrival delay reduces available turn time, and arrival delay translates into GDD when the scheduled turn time is small. Finally, the *actual turn time* is the time between actual arrival to the gate and push-back from the gate.

Neither the ASQP nor the ETMS data contain the information needed to link arrivals with departures and to determine the lengths of turn times. Some of this missing information may be inferred from the carriers' schedules. If an arrival and a departure have the same flight number, then it is likely that the two operations are performed by the same aircraft. However, at an airport where many flights terminate (such as Logan), aircraft use

different flight numbers between arrival and departure. Another approach is implemented within MITRE's National Airspace System Performance Analysis Capability (NASPAC), a simulation model used for strategic planning [13]. An "Itinerary Generator" matches scheduled arrivals with scheduled departures at each airport to find pairs of flights that may be connected as 'turns.' The matches are made according to the schedule, aircraft type, an assumed minimum turn time, and a first-come-first-served algorithm. While a good first-order approximation, the method will incorrectly match flights. Mismatches are most likely at hub airports where there are banks of arrivals and departures, and many flights are performed by the same equipment type.

Rather than trying to infer turns from published schedules, we obtained data that enabled us to identify the turns directly. We combined data from the ASQP with the Logan gate schedules of three major carriers operating at Logan airport: American Airlines and two carriers which wished to remain anonymous. The schedules of each of the three did not apply to the same time periods; they covered turns made from December 1991 to March 1993 (American), June to August 1991 ('Carrier B'), and December 1992 ('Carrier C'). The ASQP data contains the arrival delay and gate departure delay for each operation at Logan, while the gate schedules allowed us to link arrival delays with departure delays.

Figures 4-5 through 4-7 display the available turn times and GDDs of pairs of operations identified as turns for each carrier. A negative available turn time indicates that the aircraft arrived *after* its scheduled departure time. With few exceptions, a negative available turn time corresponds with a large GDD. The actual turn time is the sum of the available turn time and the gate delay. Most flights have actual turn times above 30 min., but some flights seem to have gained time on the ground (a Carrier B flight, for example, seems to have an actual turn time of -2.7 hours).

Figure 4-8 displays the cumulative percentage of flights with actual turn times within a given number of minutes. For flights operated by American Airlines, for example, 1% of the flights seem to have turned within 1 minute, 3.4% within 30 minutes, and 41% within one hour. It is interesting to note from Figure 4-8 that the three carriers differ in the speed of their quickest turns. A full 10% of Carrier B's flights turn within 30 minutes, while 5% of Carrer C's and 3% of American's do the same. The speed of a turn is influenced by the size of the aircraft, the load factor, the amount of baggage, and the size of the ground crew dedicated to cleaning and re-outfitting the aircraft. Turns may also be delayed by any

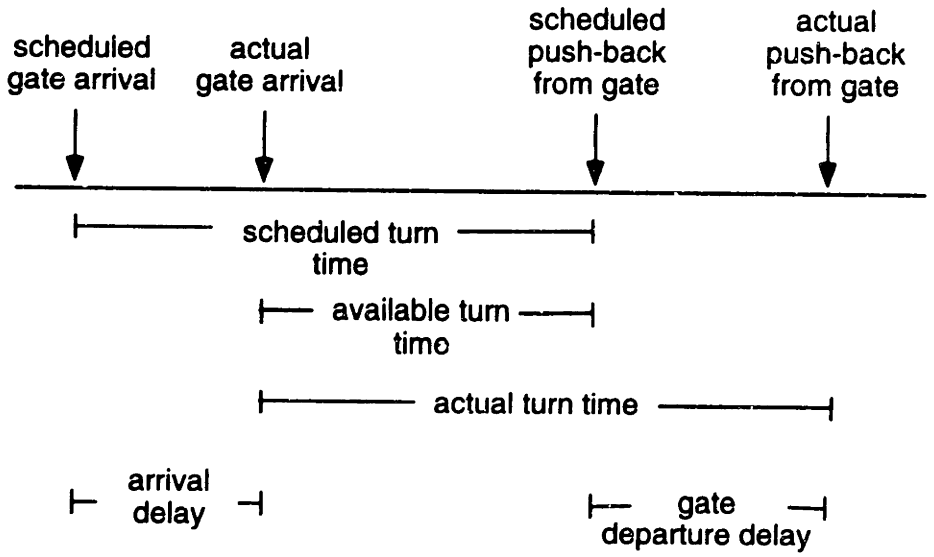


Figure 4-4: Arrival, turn, and departure of an aircraft

number of the factors we have already mentioned, such as weather, late connecting flights, late passengers, baggage, or crew.

All three carriers operated flights that seemed to have impossibly small, or negative, turn times. It is likely that these flights are beneficiaries of an aircraft exchange, or *swap*. For example, five of the seven Carrier B flights with negative turn times in Figure 4-6 were produced by the same turn. A closer look at the data indicates that this flight is the last flight of the day to a hub of the carrier, where it connects with many more flights. If the flight is late, Carrier B may exchange aircraft to keep the flight on-time. As a rule of thumb, we will assume that any flight with an actual turn time of less than 25 minutes benefited from an exchange. This included 2%, 3%, and 0.3% of American, Carrier B and Carrier C's flights, respectively. We may not have correctly identified all swaps, for it is possible that a flight with an actual turn time greater than 25 minutes is a swap while a few flights may turn without a swap in less than 25 minutes. However, the cumulative percentages of actual turns in Figure 4-8 for all three airline begins to rise between 20 and 30 minutes, and 25 minutes appears to be a reasonable demarcation between those few swaps with impossibly quick turn times and the vast majority of aircraft which arrived, turned, and departed according to the connection information in the gate schedules.

For flights that were not swaps, a simple *aircraft turn model* seems to be an accurate

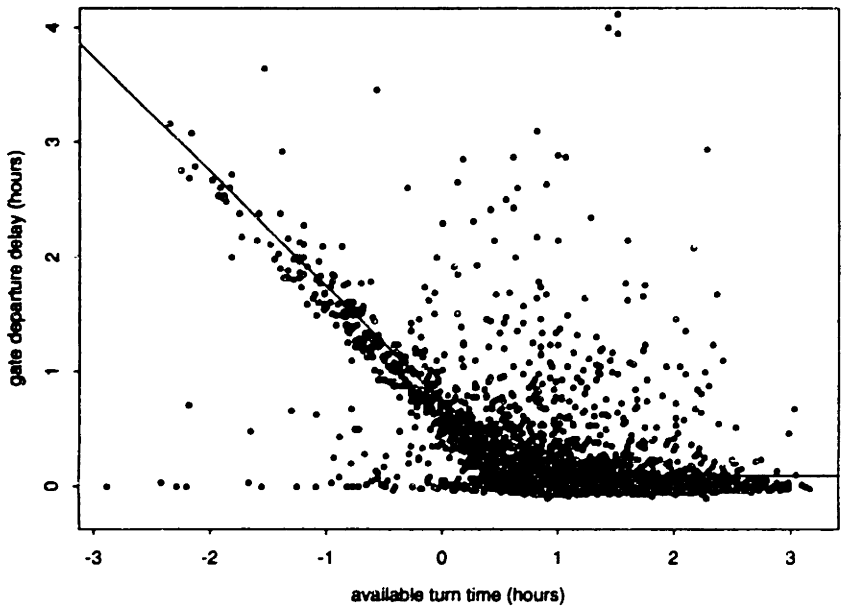


Figure 4-5: Available turn time and GDD for American Airlines flights turning at Logan (5347 flights from December, 1991 - March, 1993). The solid line is the aircraft turn model with a minimum turn time of 39 minutes.

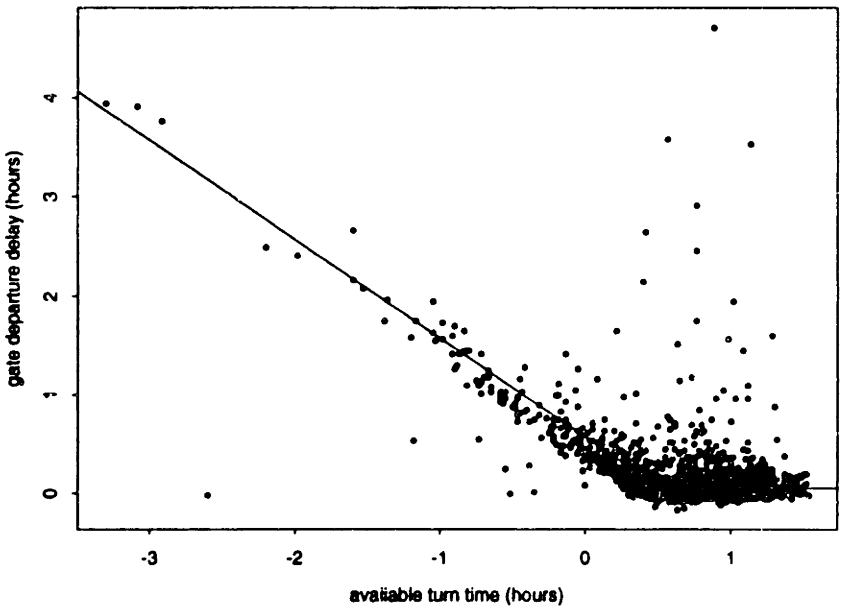


Figure 4-6: Available turn time and GDD for Carrier B flights turning at Logan (2732 flights from June - August, 1991). The solid line is the aircraft turn model with a minimum turn time of 31 minutes.

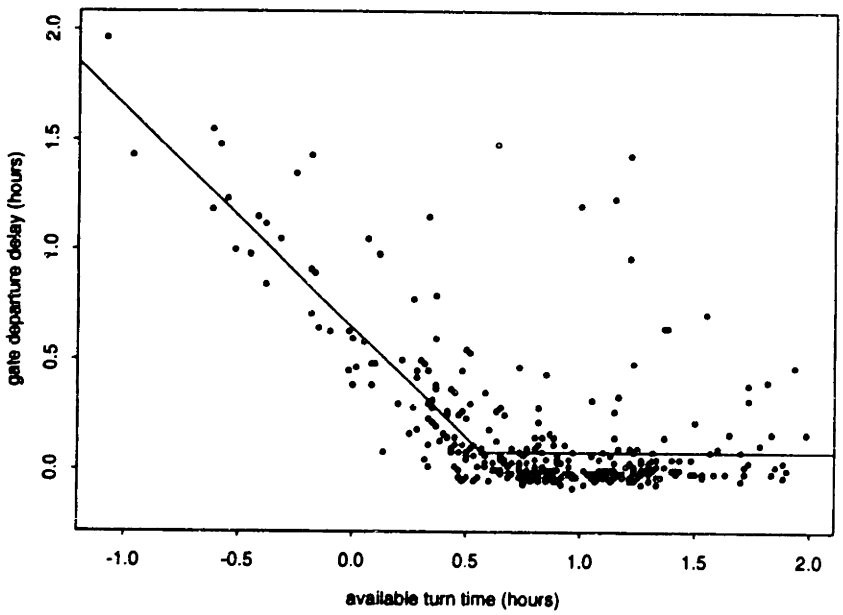


Figure 4-7: Available turn time and GDD for Carrier C flights turning at Logan (336 flights from December, 1992). The solid line is the aircraft turn model with a minimum turn time of 34 minutes.

description of GDD:

$$g_i = C + \beta\, min(0, M - a_i) + e_i \tag{4.7}$$

where

g_i = GDD of flight i (minutes)

C = GDD assessed to all flights

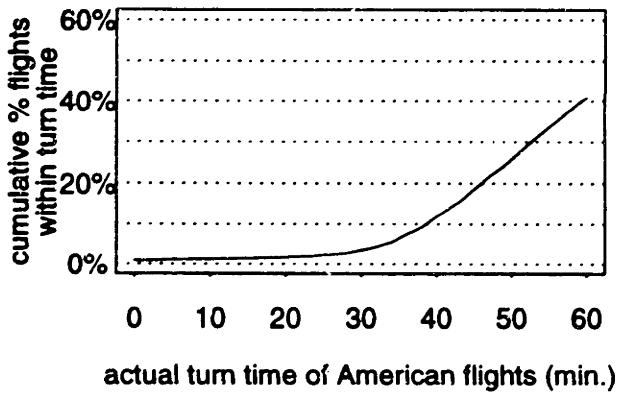
M = minimum turn time

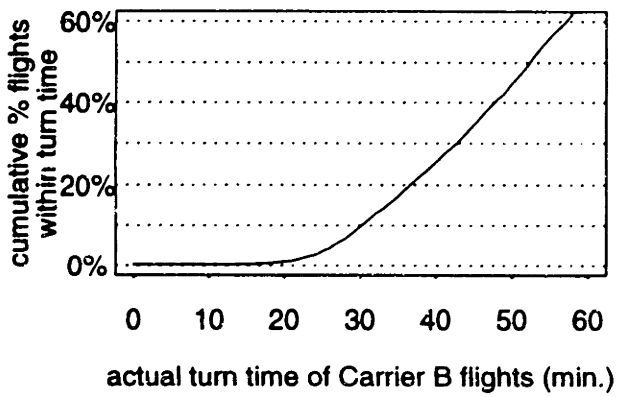
a_i = available turn time of flight i

β = additional GDD due to each minute of arrival delay once a_i falls below M

e_i = GDD not explained by the model

Given that there are no arrival delays, flights average C minutes of GDD. When the available turn time falls below some threshold value M, each minute of arrival delay translates into β minutes of departure delay.





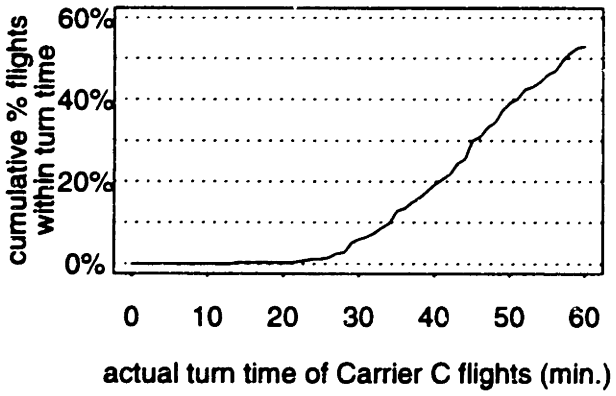


Figure 4-8: Cumulative percentage of flights turning within a given number of minutes. For example, 52% of Carrier C's flights had actual turn times that were less than or equal to 60 min.

Carrier	$\hat{M}$	$\hat{C}$ (σ_C)	$\hat{\beta}$ (σ_β)
American	39	4.9 (0.3)	1.0 (0.01)
Carrier B	31	3.4 (0.2)	1.0 (0.02)
Carrier C	34	5.5 (1.1)	1.0 (0.1)

Table 4.4: Parameter estimates and standard errors for the piecewise linear regression for the turn model. All flights with actual turn times less than 25 minutes are assumed to be swaps, and were not included in the data set used to fit the model.

	R^2 (no swaps)	R^2 (all flights)	estimated % swaps
American	0.50	0.43	2.1%
Carrier B	0.55	0.50	3.4%
Carrier C	0.40	0.40	1.2%

Table 4.5: R^2 values of the turn models evaluated with without and with flights that were assumed to be swaps. The last column displays the percentage of flights assumed to be swaps.

A procedure based on piecewise linear regression was employed to estimate parameters M, C, and β for each carrier. Flights thought to be swaps were discarded when fitting the models. The results are shown in Table 4.4 and the models are shown as solid lines in Figures 4-5 through 4-7. The iterative regression procedure took advantage of the fact that, given any value of M, the turn model may be transformed into a model which is linear in the parameters C and β. This simple transformation allows the use of ordinary least-squares regression (OLS) to find values of $\hat{C}$ and $\hat{\beta}$ which minimize the sum of squared residuals. For each carrier, we used OLS to fit models over a range of M-values from 20 min. to 60 min., and set $\hat{M}$ equal to the value which produced the model with the lowest sum of squared residuals. Given this optimal $\hat{M}$, OLS was employed to find the estimates and standard errors for C and β.

In Figure 4-8, Carrier B seems to have a higher percentage of shorter turns, and this difference is reflected in the parameters of the turn models. The three carriers have different minimum turn times, with American having the largest value of $\hat{M} = 39$ min. For all three carriers, β was close to 1.0, so that on average one minute of arrival delay produces one minute of departure delay when available turn time falls below $\hat{M}$. The differences in both $\hat{M}$ and $\hat{C}$ may be attributable to differences in the carriers' equipment and operational practices, or differences in airport conditions during the distinct time periods covered by the data for each carrier. An important factor may be the shuttle service that Carrier B operated out of Logan, for we would expect Shuttle flights to have relatively quick turns.

Tables 4.5 and 4.6 display the goodness of fit and forecast accuracy of the models generated for each carrier. Table 4.5 displays the models' R^2 values for the non-swapped flights which were used to fit the models, as well as for all flights. The models reduce the mean square forecast error by between 40 and 50%. By comparison, the factor-based linear

Carrier	MAD from model	MAD from mean	MAD % decrease	RMSD from model	RMSD from mean	RMSD % decrease
American	9.3	15.1	38%	21.3	27.4	22%
Carrier B	5.8	9.9	41%	13.9	19.7	29%
Carrier C	8.7	15.2	43%	18.3	24.1	24%

Table 4.6: Performance of models for gate departure delay. Mean absolute deviation (MAD) and root-mean squared deviation (RMSD) of delays from an aggregate mean and from the aircraft turn model. The '% decrease' refers to the percentage reduction in MAD or RMSD if the model is used instead of the aggregate mean.

model of equation 4.1 reduced the mean square error by 10%. While the two sets of models were tested with different samples of flights, the 'turn model' is undoubtedly more successful in predicting GDDs. Table 4.6 shows the mean absolute deviation (MAD) and root mean squared deviation (RMSD) of the observed GDDs from each model, as well as from the overall mean GDD of the carrier. The three turn models decrease MAD and RMSE by about 40% and 25%, respectively, compared to a constant forecast of the overall mean. The turn model is particularly successful in predicting the largest GDDs. Figure 4-9 compares the predictive accuracy of the Carrier B turn model with the predictive accuracy of the overall average GDD for subsamples of progressively larger GDD. The mean absolute error is displayed for all flights and for flights with over 30 and 60 minute GDDs. As the length of the delay grows, the advantages of the turn model becomes more pronounced.

4.4 Conclusions and Extensions

In this chapter we have developed models of GDD based on environmental factors, such as weather conditions, as well as models based on the arrival delays and gate turns of flights. The models based on environmental factors produced only marginal gains in GDD forecasting accuracy. Previous studies of delay codes of a major airline found that large percentages of delays may be attributed to mechanical problems, aircraft crew performing last-minute duties, or passenger delays. Many of these are difficult to predict, and it is not surprising that the environmental factor model did not significantly decrease the variance of GDD forecasts.

On the other hand, a simple turn model using arrival delays to predict GDD produced

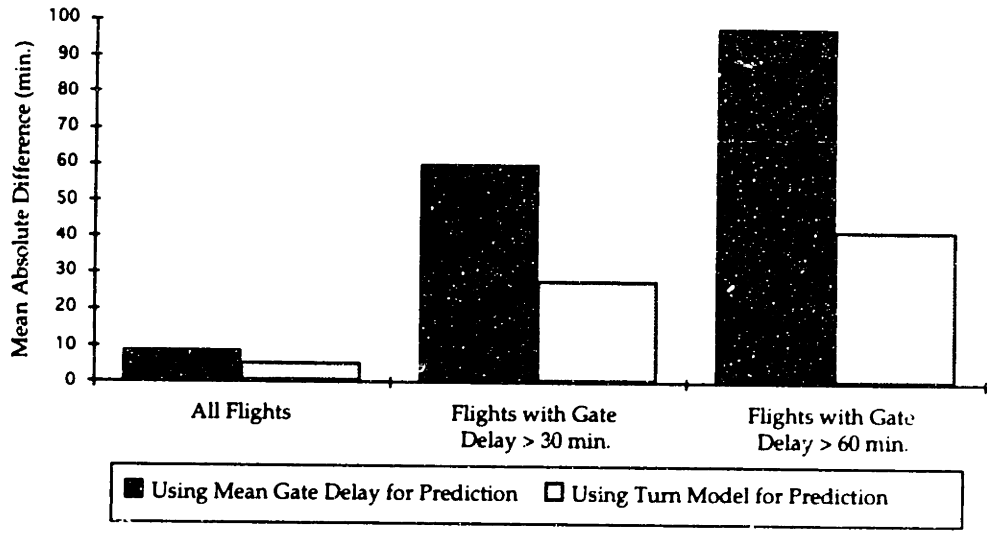


Figure 4-9: Mean absolute deviation between observed GDDs, the turn model, and the overall mean GDD for all Carrier B flights, as well as flights with actual GDDs above 30 and 60 minutes.

much larger improvements. This is especially encouraging if one wishes to produce predictions with a forecast horizon of one to three hours. ETMS predictions of aircraft trajectories are reasonably accurate once an aircraft is in the air, so accurate arrival delay forecasts for many flights are available over one hour before arrival and subsequent departure from the downstream airport. The aircraft turn model should be effective for real-time delay prediction.

The results of this chapter also suggest further research on aircraft turn models. The differences between estimated values of the minimum turn time, M, in Table 4.4 demonstrate that different carriers have, on average, different minimum turn times. This may be explained by differences in fleet composition, scheduling practices, staffing levels, or operational procedures. In general, the models and the supporting data offer opportunities to infer information about the competitive practices of the carriers. In addition, improved piecewise linear models might distinguish between equipment types or individual flight numbers to provide a better fit to the data.

Table 4.5 displays the percentage of each carriers' flights that were classified as a 'swap'. Although small in number, these flights often fall far from the turn model. The model will be much less effective at airports where swaps, substitutions with spare aircraft, or aircraft exchanges for maintenance are common. In particular, carriers at their hub airports may have more opportunities to perform swaps and may be more likely to exchange turning

aircraft in order to minimize the effect of arrival delays. One further area of research might be the construction of statistical models which track previous information about aircraft swaps and predict future exchanges.

However, carrier real-time operations are extremely dynamic, and carrier strategies change from day to day and month to month. The FAA should continue to pursue opportunities for real-time data exchange with the carriers, for any model of turns and swaps constructed without carrier data is likely to produce forecasts which are inferior to the carriers' own predictions. The aircraft turn models demonstrate the large improvements in predictive accuracy that are possible with access to only a small amount of carrier data, the gate schedules. More cooperation between the carriers and the FAA may produce even larger benefits.

Chapter 5

Aircraft Flow Models and Capacity Estimates

In order to predict take-off times of aircraft, we are developing methods for predicting both gate departure delays and roll-out times (the time between gate push-back and take-off). In general, the time to get from the gate to the runway is likely to increase as congestion on the airfield increases. For example, taxiing aircraft may be blocked by other aircraft on the apron or may spend time in a queue at the head of the departure runway. This chapter develops models of an airport's departure process which describe aircraft movements as a deterministic flow from the gates to the departure runway and into the air. The rate of flow onto the airfield is governed by the rate of aircraft push-backs; the rate of flow out of the airfield is limited by airport capacity.

A build-up of aircraft on the airfield is likely to produce an increase in the roll-out time of individual aircraft, and predictions of airfield congestion should be useful for predicting the roll-out times of individual aircraft. However, these models and forecasts may also be of special interest to air traffic controllers and the carriers, apart from their value for predicting individual aircraft roll-out times. For example, they may allow tower controllers and carrier gate managers to plan for, and perhaps prevent, ground congestion by holding aircraft at the gate or sending them to less congested runways.

The next section of this chapter provides a brief description of related research on airport capacity and queueing delays. Section 5.2 motivates our models by presenting ETMS data on airport congestion and take-off rates. Section 5.3 describes the aggregate flow models

while Section 5.4 describes how the data are used to produce real-time estimates of take-off capacity. In both sections, simple examples demonstrate how the models are fitted to data and generate forecasts of airfield congestion. Section 5.5 presents empirical tests with data from Logan Airport. It compares the accuracy of the proposed models and tests the sensitivity of the models to the underlying assumptions about the departure system. The final section suggests areas for further work.

5.1 Previous Research

A large amount of research on airport capacity and delays to aircraft at airports has been produced by the FAA, government consultants, and academics. An article by Odoni [27] provides an overview of early and recent published research. Representative early work includes articles by Blumstein [8] and Newell [26]. Blumstein's simple but powerful model for aircraft arrivals illustrates the relationship between the maximum landing rate and factors such as separation requirements, length of the glide path, and aircraft velocity. Newell extends Blumstein's work by modeling the interaction between arriving and departing aircraft. He derives *convex capacity curves*, which show the maximum rates of arrivals and departures as the arrival/departure mix is varied. Odoni describes more recent work on delay analysis for congested airports. Airports are modeled as queueing systems with time-varying demand and service rates. In terms of our own aircraft departure problem, the rate of push-backs from the gate may vary with time, and may sometimes exceed the maximum take-off rate, so that $\rho > 1$ temporarily. Such a system cannot be adequately described using more traditional homogeneous, steady-state queueing analysis.

These analytical delay models describe the expected system behavior, given a probabilistic description of the demand and service processes. In order to simplify the analysis, the probabilistic descriptions belong to a family of general stochastic processes. For example, arrivals to the system may be Poisson and service times distributed as Erlang random variables. The output of the model may be the expected value of the queue lengths or the probability distribution of the number in the queue. These predictions of average behavior are most useful for strategic planning, but are less useful for predicting specific delays to actual flights. The models provide a useful guide to the factors which influence airport capacity and delays but are not directly applicable to our problem of real-time delay pre-

diction. For example, if the ETMS supplies accurate predictions of the push-back times of individual aircraft, a description of demand with a general stochastic process, such as the Poisson, would ignore this detailed information. In addition, the arrival of real-time data also creates the opportunity to update the model, given observed airport conditions. This capability does not exist in the more traditional planning models.

Empirical studies by Gilbo [19] and a study by MITRE [12] are closer in spirit to the work of this chapter, and more detailed descriptions of these studies will be used to motivate our approach. Both studies produce estimates of airport capacity based on counts of aircraft operations (arrivals and departures). The number of operations in discrete intervals is tabulated, and capacity is estimated from the maximums of these observed counts after adjusting for the presence of outliers. The MITRE study is applied to arrivals at Chicago O'Hare, while Gilbo's paper derives estimates of the convex capacity curves at each of the largest and busiest domestic airports.

While a reasonable approach, the use of observed maximums has a few potential pitfalls. First, these estimates underestimate capacity in unsaturated systems, for capacity will only be observed when the system is saturated. For example, suppose that there are two runway configurations with identical capacities and that during the period of data collection the first configuration has demand which saturates the system while the second has low demand. A capacity estimate calculated from the observed maximum number of operations may produce a reasonable estimate for the first configuration but will underestimate the capacity of the second. The empirical studies described above are based exclusively on departure counts, and no attempt was made to determine whether, or for how long, the systems were under stress. However, information is now available that will allow us to keep track of the number of aircraft in the process of rolling out. This will serve as a rough estimate of system demand and will allow the flow models introduced in this chapter to distinguish between saturated and unsaturated conditions.

In addition, the observed number of observations will fluctuate over time, even if there is sufficient demand to maintain operations at capacity. In Gilbo's paper, the capacity curves are described as percentiles of the arrival and departure counts over fifteen-minute intervals. Distinct curves were found for the 90th, 95th and 99.5th percentiles. Gilbo points out that the higher percentiles may overestimate capacity since the observed fifteen-minute rates could not be sustained for a longer period of time. Newell recognizes this possibility

when he defines capacity to be:

> a maximum average flow that a facility can accommodate over a time period long enough to include a large count (say 100 or more) and which could, in principle, be sustained for an infinitely long time (if one had an arbitrarily large reservoir of aircraft) ([26], p. 209).

For aircraft arrivals and departures, take-off rate fluctuations are often caused by temporary changes in the aircraft mix. When a Beechcraft with a capacity of 14 passengers follows a DC-10 with a capacity of 380, the Beechcraft must allow a large headway for safety, and the service rate is temporarily reduced. Many small (or even large) aircraft in a row increase the observed rate of operations. Given an overall traffic mix, however, one might consider these fluctuations to be variations from the overall capacity for that traffic mix. We will adopt Newell's definition of a 'true' of 'sustainable' take-off rate as our definition of capacity.

While some deviations from capacity are short-term fluctuations, there may also be more systematic changes in capacity over time. These changes may be caused by changes in weather, runway configuration, or controller practices. Estimates of capacity based on maximum counts of operations ignore the time-dependent nature of capacity, for all time periods are given equal weight. A capacity estimate at 8:00 may be improved by giving more weight to counts of operations at 7:50 than to counts from the previous day.

Current efforts at Atlanta Hartsfield International Airport have been directed towards this area of dynamic queueing and capacity estimation [2]. A system for real-time data exchange has been developed by the Atlanta Tower and the airport's primary carriers. The existing system is a prototype which allows both Tower controllers and the carriers' ramp controllers to share scheduled and actual push-back and arrival times. Such a system would be an ideal platform on which to build the flow models developed in this chapter.

5.2 Empirical Description of Airfield Congestion and Take-off Rates

Before formulating models for aircraft roll-outs, it is useful to examine the ETMS data to see the evolution of airport congestion over time and to investigate the relationship between congestion and the take-off rate. We focus our attention on $N(t)$, the number of aircraft

rolling out at time t. Given a more complete data set, $N(t)$ would be easy to calculate:

$$N(t) \quad = \quad \# \text{ push-backs by time } t - \# \text{ take-offs by time } t \tag{5.1}$$

Unfortunately, $N(t)$ cannot be calculated directly in this way from the ETMS data. Recall from Chapter 3 that 56% of departures are *unmatched* flights without known push-back times. Push-back times are known for the 44% of flights which did match between two crucial databases. Section 5.2.1 below describes procedures to estimate $N(t)$ using the available data from all flights. The procedures themselves are rather technical and are not of great methodological interest, but the resulting estimates of $N(t)$ will be important for motivating and testing the aggregate flow models. In section 5.2.2 we use these ETMS estimates of $N(t)$ to examine empirically the relationship between congestion and take-off rates. Patterns in these data will motivate the flow models.

5.2.1 Inferring the Number Rolling Out from Available Data

For the 56% of flights which are unmatched, push-back times are not included in the data set, but scheduled departure times and actual take-off times are known. Two methods are proposed for assigning push-back times to unmatched flights. The first calculates forward from the flight's scheduled departure time by adding an average gate departure delay which is calculated from the matched flights. The second calculates backwards from the flight's actual take-off time by subtracting an average roll-out time.

For the first method, let SCH(i) be unmatched flight i's scheduled departure time, according to the filed flight plan. Let $\bar{g}$(SCH(i)) be the *average* gate departure delay of matched flights with scheduled departure times within a ten-minute window around SCH(i). The first estimated push-back time is PB1(i) = SCH(i) + $\bar{g}$(SCH(i)) (see figure 5-1).

The second method calculates push-back time by subtracting a roll-out time from flight i's actual take-off time, TO(i). Let $\bar{r}$(TO(i)) be the average roll-out time for matched flights with take-off times within a ten-minute window around TO(i). The second estimated push-back time is PB2(i) = TO(i) - $\bar{r}$(TO(i)).

Given either PB1 or PB2, the number rolling out, ($N(t)$), can be calculated from known take-off times TO(i) and estimated push-back times using equation 5.1. Figure 5-2 shows

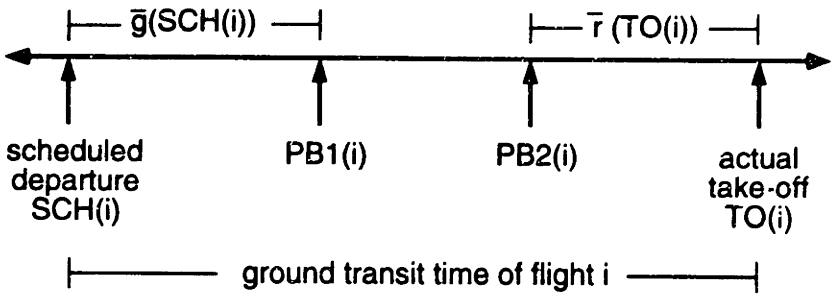


Figure 5-1: Push-back time of unmatched flight i estimated from the scheduled departure ($PB1(i)$) and from the actual take-off ($PB2(i)$).

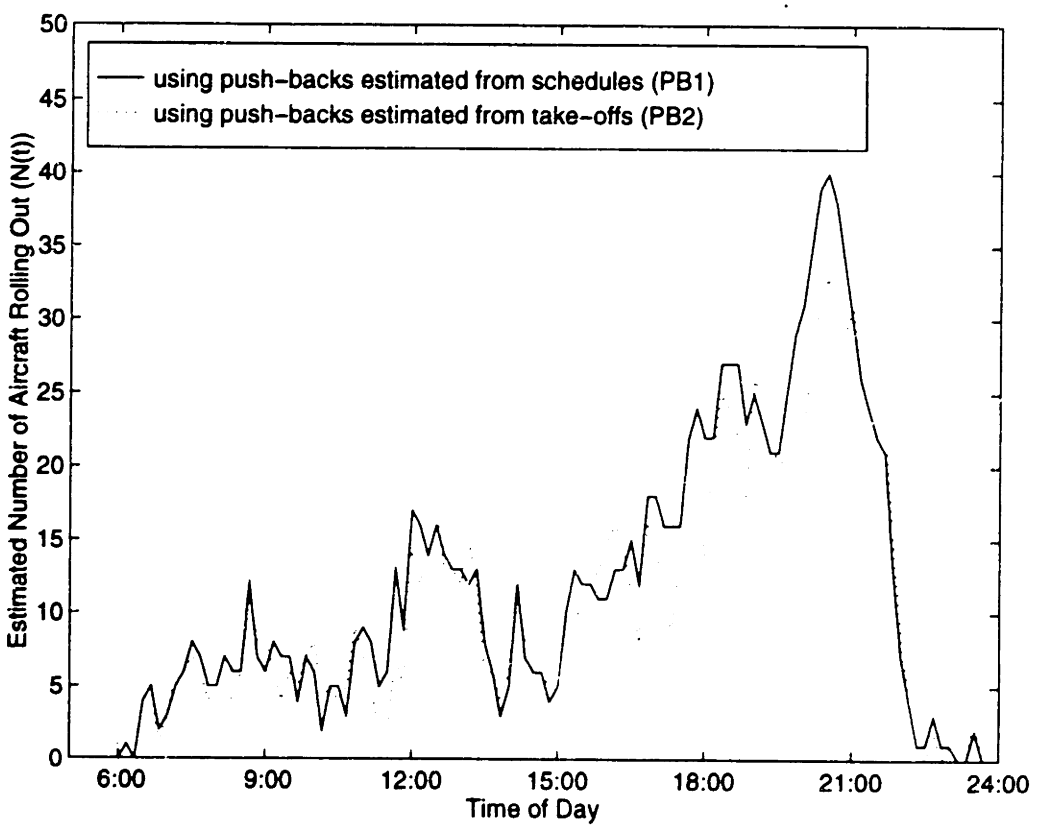


Figure 5-2: Estimates of the number of aircraft rolling out ($N(t)$) on August 18, 1991 using push-backs estimated by adding gate delay to the scheduled departure (PB1(i)) and subtracting roll-out time from observed take-off (PB2(i))

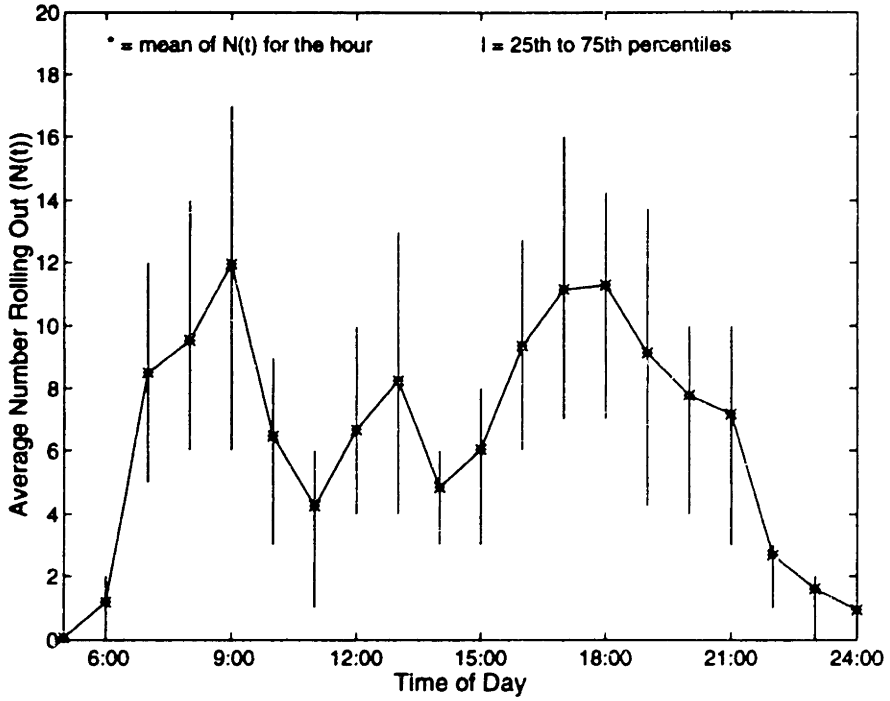


Figure 5-3: Hourly average of $N(t)$, the estimated number of aircraft rolling out. Averages are taken over a 27-day sample from March and August, 1991.

estimates of $N(t)$ on August 18, 1991.[1] The number of roll-outs calculated from the two methods sometimes grows far apart in IFR or LIFR. This may be explained by the effect of poor weather on general aviation and commuter flights, which make up the bulk of the unmatched data. In the worst conditions, these flights tend to suffer longer delays than flights operated by the large carriers. If these delays cause an unmatched flight i to have a long ground transit time, then PB1(i) will fall relatively early in the flight's GTT, and the flight is assumed to have a lengthy roll-out time. This flight will be included in the count for $N(t)$ over many periods. On the other hand PB2(i) will place the push-back later in the ground transit time, under the assumption that a bulk of the long delay was taken at the gate. This removes the aircraft from the airfield, and from the count of $N(t)$, during the long estimated time at the gate. The most extreme example from our data set occurred on August 20, 1991, the day after the hurricane passed New England. At one point, the value of $N(t)$ calculated from PB1(i) was 25 while the value calculated from PB2(i) was 5. The PB1 push-back estimates placed 20 more aircraft in the roll-out stage; PB2 placed these aircraft at the gate. Averaged over our 27 day sample, however, the absolute value

[1]By the evening of August 18, Hurricane Bob had reached North Carolina. Figure 5-2 displays unusually large values of $N(t)$. For example, compare it with Figure 5-3.

of the difference between these two $N(t)$ estimates is only 0.5 aircraft. This is encouraging, for the values of $N(t)$ calculated from PB1 and PB2 are rough upper and lower bounds on the number of aircraft rolling out. On average, either estimate should be close to the true value.

There are many alternatives to PB1 and PB2, such as a weighted average of the two or a probabilistic model for each push-back. However, we have made the reasonable decision to adopt PB2(i) as the push-back time of each unmatched flights. Roll-out times of matched flights should be good estimates of roll-out times of unmatched flights, for both groups face similar conditions on the airfield. In addition, gate departure delays have a higher variance and a much lower correlation between flights, making PB1 unreliable. In fact, the scheduled departure times SCH(i) used to calculate PB1 have different interpretations for different types of flights. The scheduled departure time in a filed flight plan of an executive jet is often a *proposed take-off time* and not a scheduled push-back time. Adding the average carrier gate departure delay to the proposed take-off time of these flights makes little sense. Push-back estimates PB2, on the other hand, depend only on the relatively reliable take-off times. Therefore, for the remainder of this thesis, the quantity '$N(t)$' will represent the estimate of the number rolling out calculated from push-back times PB2.

In our sample, we find that the minimum value of $N(t)$ was zero (i.e. at 5 A.M. on most days), while the maximum number was 33 (during the evening of August 18, as shown in Figure 5-2). The hourly average values of $N(t)$ at Logan are shown in Figure 5-3, along with the range from the first quartile to the third quartile. Note the distinct morning and evening peaks, attributable to the increase in take-off demand during these periods. In the next section we will investigate how departure rate varies with $N(t)$.

5.2.2 Observed Relationship Between Airfield Congestion and Take-off Rates

One would expect that during periods when $N(t)$ is small, there would be fewer take-offs. As $N(t)$ rises and departure demand increases, the observed number of take-offs should rise until take-off capacity is reached. Here we examine empirically the relationship between $N(t)$ and the take-off rate.

To produce estimates of take-off rates, we divide each day into ten-minute periods. Let

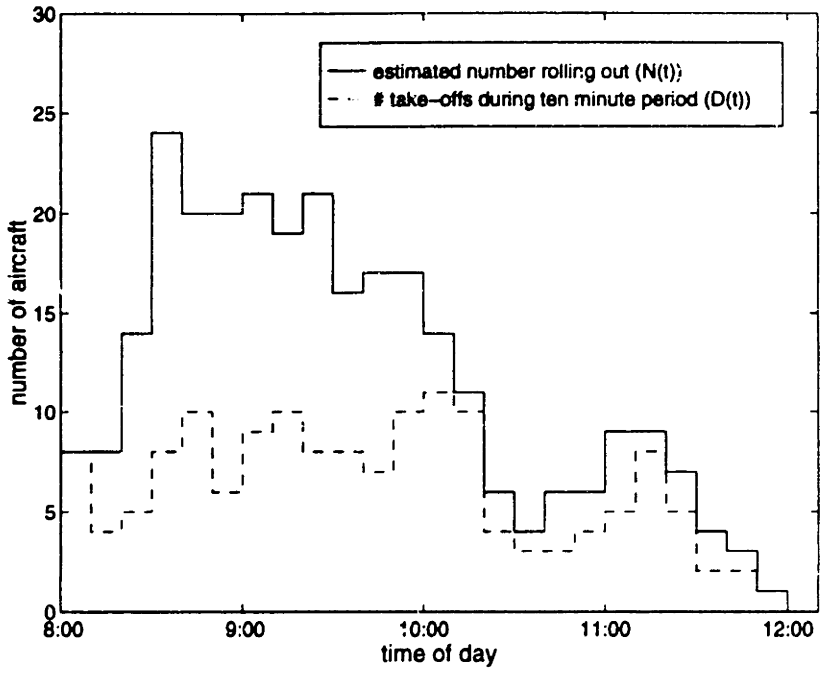


Figure 5-4: Estimated number of aircraft rolling out ($N(t)$) at the beginning of ten minute periods and the number of observed take-offs ($D(t)$) during each period.

t be the time period's index, and let

$$N(t) = \#\text{ of aircraft rolling out at beginning of period } t \tag{5.2}$$

$$D(t) = \#\text{ of take-offs during time period } t. \tag{5.3}$$

$$B(t) = \#\text{ of push-backs during time period } t$$

Note that $N(t+1)$ can be calculated from the recursive 'conservation of aircraft' equation:

$$N(t+1) = N(t) - D(t) + B(t) \tag{5.4}$$

Figure 5-4 shows estimated values of $N(t)$ and observed take-off counts $D(t)$ from 8:00 am to noon on August 27, 1991. The largest rise in $N(t)$, at 8:40, was caused by sixteen push-backs during the preceding ten-minute period. The observed number of take-offs rises with $N(t)$, and seems to have an upper limit around $8-10$ aircraft per ten minute period (a rate of $48-60$ aircraft/hour). It is possible for $D(t)$ to exceed $N(t)$, for aircraft may push-back and take-off within a single ten minute period, although this does not happen during any of the periods in Figure 5-4.

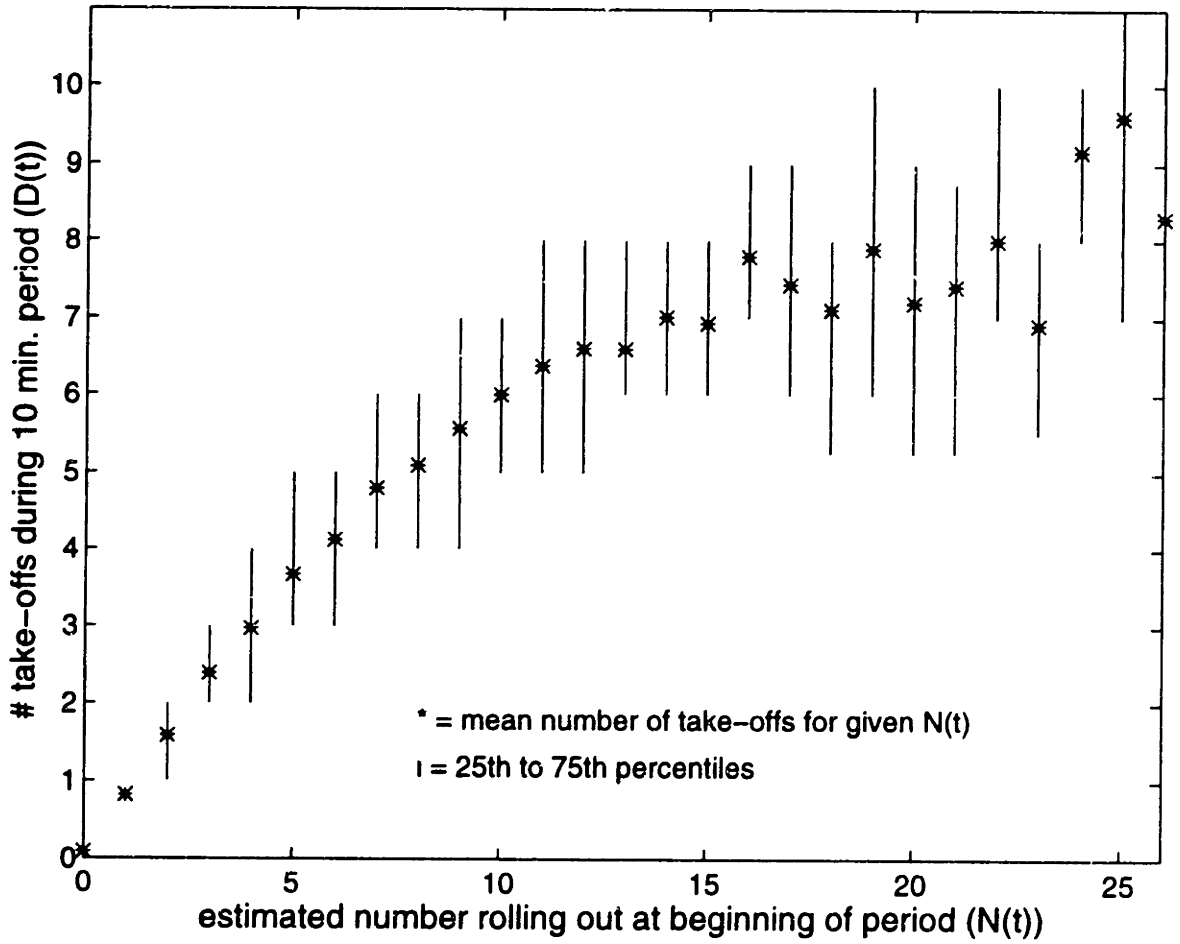


Figure 5-5: Observed number of take-offs during each period ($D(t)$)) rises with the estimated number of aircraft rolling out at the beginning of each time period ($N(t)$). An '*' indicates the mean number of take-offs while the vertical bar is the range from the first to the third quartile.

Values of $B(t)$, $D(t)$, and $N(t)$ were calculated for 3,105 ten-minute periods over the 27 days in our data set. Figure 5-5 displays the mean number of take-offs $D(t)$ per period for each value of $N(t)$. The vertical bars describe the range from the first to the third quartile. The mean take-off rate rises with $N(t)$ and levels off between seven and eight aircraft per ten-minute period. This upper plateau might be used as a rough estimate for take-off capacity, but it is important to remember that Figure 5-5 was generated from data over a time period which exhibited a wide variety of weather and runway conditions. The interquartile ranges shown in the figure demonstrate how variable capacity may be. The distance between the first and third quartiles of take-off counts during the ten-minute period is as large as five aircraft. Some of this variation may be caused by random fluctuations around the capacity, and some by changes in the capacity itself. The shape of Figure 5-5 motivates the following models of aircraft flow.

5.3 Deterministic Models for Departures

The relationship between airfield congestion and take-off rates displayed in Figure 5-5 is not surprising when one considers the dynamics of the departure process. Most of the $N(t)$ aircraft are moving towards the departure runway, and during each ten-minute period some proportion of them join the departure queue or are able to take-off. These observations are the basis of the following two deterministic models for departures. Each model is specified by three sets of equations: *system equations* which describe the underlying dynamics, *forecast equations* which generate forecasts based on these dynamics, and *update equations* which correct estimates of $N(t)$, given observed data.

5.3.1 One-stage Deterministic Model

At major airports, departing aircraft push-back from gates and taxi to one of a few possible departure runways. At times, there may be more than one active departure runway. However, in this model we will consider all active runways to be a single 'server' with a capacity equal to the take-off capacity of all active runways. This is likely to be a reasonable approximation for Logan, where there is usually one primary departure runway and operations on runways often interact. This approximation may not be reasonable for an airport with more than one independent, major departure runway, such as Atlanta Hartsfield or Denver

International.

During a T-minute time period, the number of aircraft that enter the system is equal to the number of push-backs, $B(t)$. After push-back, each aircraft begins its roll-out to the departure runway. The one-stage deterministic model assumes that if the aircraft is allowed to proceed unobstructed, it will reach the departure runway $\bar{R}$ minutes after push-back. However, rather than keeping track of individual aircraft, the model assumes that taxiing aircraft are distributed uniformly along their $\bar{R}$-minute taxi-times. If there are $N(t)$ aircraft taxiing at time t, then all $N(t)$ aircraft will have reached the departure runway $\bar{R}$ minutes later. According to the model, the number that reach the departure runway in time T is $(T/\bar{R})N(t)$ for $T \leq \bar{R}$. Note that this is the expected value of the number which reach the runway in time T under the assumption that all $N(t)$ aircraft are distributed uniformly along the taxiway. This can also be viewed as a 'fluid' model, where each aircraft flows from gate to runway at a rate $1/\bar{R}$ aircraft per minute. If only one aircraft is taxiing, a fractional aircraft, $T/\bar{R}$, reaches the runway during T minutes.

The model also assumes that the departure runway has a finite capacity of $\mu(t)$ aircraft/minute. The runway may not be able to fulfill demand for take-offs when $(T/\bar{R})N(t) > T\mu(t)$. Therefore, the number of departures during period t, $D(t)$, is the minimum of the demand from aircraft rolling out and the capacity. The following difference equation describes this system:

System Equation

$$N(t+1) = N(t) - D(t) + B(t) \tag{5.5}$$

$$D(t) = \min[\frac{T}{\bar{R}}N(t), T\mu(t)] \tag{5.6}$$

where the definitions of $N(t)$, $D(t)$ and $B(t)$ can be found above in (5.2).

The following example demonstrates the evolution of $N(t)$ under the one-stage model.

EXAMPLE: Let the average roll-out time $\bar{R}$ = 15 min., the period length T = 10 min., and the capacity $T\mu(t)$ = 8 aircraft per ten-minute period for all t. Table 5.1 and Figure 5-6 describe a simple example of the evolution of this one-stage model in which push-backs only occur during the first two periods: $B(1) = 9, B(2) = 18$, and $B(t) = 0$ for $t > 2$. We begin with an empty airfield ($N(1) = 0$). During the first period, 9 aircraft push-back, but since there were no

t	$N(t)$	$D(t)$	$B(t)$
1	0	0	9
2	9	6	18
3	21	8	0
4	13	8	0
5	5	3.3	0
6	1.7	1.1	0
7	0.6	0.4	0
8	0.2	0.1	0

Table 5.1: Example of the one-stage model with $T/\bar{R} = 2/3$ and $\mu(t) = 8$ aircraft per period.

aircraft rolling out at the beginning of the period, there are no take-offs. During the second period, $(T/\bar{R})N(2) = (2/3)(9) = 6$, so that six of the nine rolling aircraft reach the departure runway and take off. Meanwhile, 18 more aircraft push back. By the start of period 3, $N(t) = 21$ and $(2/3)(21) = 14$ aircraft arrive to the departure runway for take-off, so that demand exceeds capacity. The maximum of eight aircraft depart, and the 13 remaining aircraft are left uniformly distributed along the taxiway. In period 5 the system is no longer saturated, and $(2/3)5 = 3.3$ aircraft take-off. $N(t)$ continues to decrease for the remaining periods. In Figure 5-6, $D(t)$ reaches a plateau at $\mu(t)$ during periods 3 and 4 and is proportional to $N(t)$ during the remaining periods.

The system equations describe the dynamics of the model, but our goal is to forecast values of $N(t)$. While we may be able to predict $B(t)$ with some accuracy, take-off rates $D(t)$ are not known in advance. We introduce some additional notation for these forecasts:

$$\begin{aligned} \hat{N}(t|t) &= \text{estimated number of aircraft rolling out at start of time period } t, \\ &\quad \text{given all information available at the beginning of time period } t \\ \hat{N}(t+1|t) &= \text{estimated number of aircraft rolling out at start of time period } t+1, \\ &\quad \textit{given} \text{ all information available at the beginning of time period } t \end{aligned}$$

For our purposes, "all information available at the beginning of time period t" includes all observed push-backs and take-offs up to, but not including, those that occur during period t. The quantities $\hat{D}(t|t)$, $\hat{D}(t+1|t)$, $\hat{B}(t|t)$, and $\hat{E}(t+1|t)$ are defined in a similar way. For now, assume that the estimates of take-off capacity, $\hat{\mu}(t|t)$, are determined in advance from

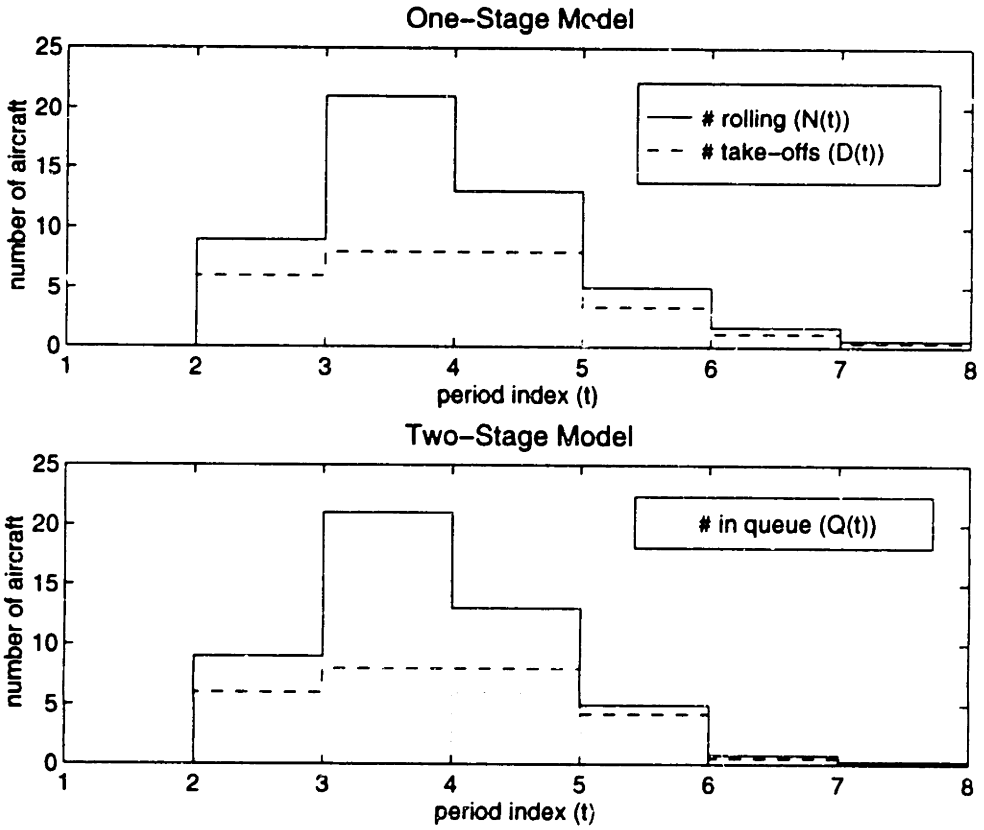


Figure 5-6: Example of the departure system according to the one-stage and two-stage models. The two differ slightly after period 4, when the departure queue in the two-stage model delivers a few more aircraft for take-off than the one-stage model.

t	$N(t)$	$D(t)$	$B(t)$	$\hat{N}(t\|t)$	$\hat{D}(t\|t)$	$\hat{B}(t\|t)$	$\hat{N}(t+1\|t)$
1	0	0	9	0	0	9	9
2	9	3	18	9	6	18	21
3	24	9	0	24	8	0	16
4	15	9	0	15	8	0	7
5	6	5	0	6	4	0	2
6	1	1	0	1	$\frac{2}{3}$	0	$\frac{1}{3}$
7	0	0	0	0	0	0	0

Table 5.2: Updated estimates and one-period forecasts using the one-stage model.

the predicted runway configuration and weather conditions. In the next section we will consider methods for estimating $\mu(t)$ from the real-time data. At first, we will also assume that predictions of push-back counts, $\hat{B}(t+1|t)$, are always correct (i.e. $\hat{B}(t+1|t) = B(t+1)$).

The forecast equation is derived directly from the system equation:

Forecast Equation

$$\begin{aligned} \hat{N}(t+1|t) &= \hat{N}(t|t) - \hat{D}(t|t) + \hat{B}(t|t) \\ \hat{D}(t|t) &= \min[\frac{T}{\bar{R}}\hat{N}(t|t), T\hat{\mu}(t|t)] \end{aligned} \tag{5.7}$$

The forecast for each period is determined by assuming a deterministic flow from the previous period, just as we did at each step in Table 5.1. These are one-step forecasts (10 minutes in advance). Additional forecasts are calculated recursively.

Finally, the *update equation* incorporates the most recent observations about push-backs and take-offs into the estimate of $N(t)$. The update equation for this model is rather obvious: the current best estimate of $N(t)$ is $N(t)$ itself.

Update Equation

$$\begin{aligned} \hat{N}(t|t) &= N(t) \\ &= \hat{N}(t-1|t-1) - D(t-1) + B(t-1) \end{aligned} \tag{5.8}$$

Now consider a variation of the simple example introduced above.

EXAMPLE: Table 5.2 shows the one-step forecasts and updated estimates for the simple demand profile introduced earlier. In this example, the actual take-off counts $D(t)$ differ from those that

are predicted by the one-stage flow model. The model predicts 6 take-offs during the second period ($\hat{D}(2|2) = 6$) while only three actually depart during the period. Since $N(2) = 9$ and there will be 18 push-backs during the second period, the model predicts that there will be $\hat{N}(3|2) = 9 - 6 + 18 = 21$ aircraft rolling out at the beginning of the third period. By the beginning of the third period, however, the system has observed that the actual number of departures during the second period was 3, so that the updated estimate of $N(t)$ is set to the correct value: $\hat{N}(3|3) = N(3) = 24$. This cycle of one-step predictions and corrections continues until there are no aircraft left on the field.

One of the fundamental assumptions of the one-stage model is that aircraft are distributed uniformly along their $\bar{R}$-minute "taxi trajectory." When the airport is saturated, however, a departure queue may build near the departure runway. The following extension to the model captures this effect.

5.3.2 Two-stage Deterministic Model

The two-stage model also describes departing aircraft with a fluid approximation, but allows aircraft to accumulate in a departure queue when the system is saturated. As in the one-stage model, the rate of flow to the departure queue is proportional to the number of aircraft rolling out. Aircraft which reach the departure queue are served at a rate $\mu(t)$ aircraft/minute. If aircraft flow to the departure runway at a rate higher than $\mu(t)$, a queue builds in front of the departure runway. One sample trajectory of this system is shown in figure 5-7. The figure displays the cumulative number of push-backs, aircraft that reach the departure queue, and take-offs. Now there are two state variables to keep track of: $X(t)$, the number of aircraft taxiing out and $Q(t)$, the number of aircraft in the departure queue. Their sum is the total number of aircraft rolling out: $N(t) = X(t) + Q(t)$. By time t in the figure, the take-off rate has fallen below demand to the runway, and a departure queue of size $Q(t)$ has developed.

Again, all aircraft flow from the gate to the departure queue in a constant taxi time, $\bar{R}$. Note that $\bar{R}$ is not the roll-out time of the aircraft, for roll-out time includes time spent in the departure queue as well as on the taxiway. Actual roll-out time is at least, and usually greater, than $\bar{R}$. The uniformity assumption that was applied to $N(t)$ in the one-stage model is assumed for the $X(t)$ taxiing aircraft in the two-stage model. Given $X(t)$ aircraft on the taxiway, then the number of aircraft that reach the departure queue in time T is

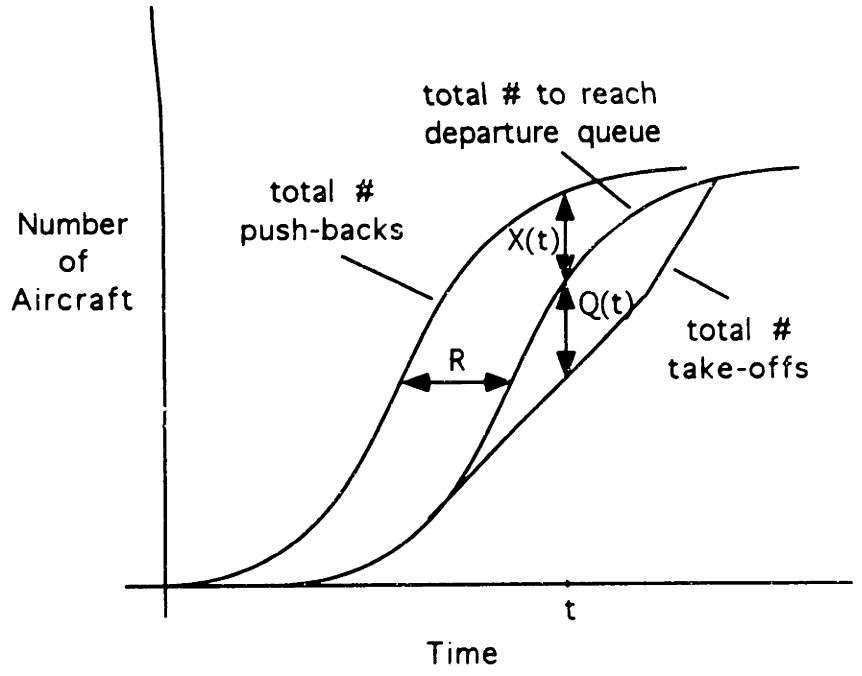


Figure 5-7: Two-stage deterministic departure process model with taxi-out time $\bar{R}$. $X(t)$ taxiing aircraft, and $Q(t)$ aircraft in the departure queue

$(T/\bar{R})X(t)$ if $T < \bar{R}$. These aircraft are added to the departure queue, which acts as a buffer for unserved aircraft. The departure demand during each time period is the sum of the queue size at the beginning of the period, $Q(t)$, and the aircraft flowing to the queue during the period, $(T/\bar{R})X(t)$.

The two-stage model is specified by the following system equations:

$$
\begin{aligned}
X(t+1) &= X(t) - Y(t) + B(t) \qquad (5.9)\\
Q(t+1) &= Q(t) + Y(t) - D(t)\\
Y(t) &= \frac{T}{\bar{R}} X(t)\\
D(t) &= \min[Q(t) + Y(t), T\mu(t)]
\end{aligned}
$$

where

$X(t)$ = number of aircraft taxiing at start of time period t

$Q(t)$ = number of aircraft in departure queue at start of time period t

$Y(t)$ = number of aircraft moving from taxi to the departure queue during time period t

$B(t)$ = number of aircraft pushing back during time period t

t	$X(t)$	$Q(t)$	$N(t)$	$Y(t)$	$D(t)$	$B(t)$
1	0	0	0	0	0	9
2	9	0	9	6	6	18
3	21	0	21	14	8	0
4	7	6	13	4.7	8	0
5	2.3	2.7	5	1.5	4.2	0
6	0.8	0	0.8	0.5	0.5	0
7	0.3	0	0.3	0.2	0.2	0
8	0.1	0	0.1	0.1	0.1	0

Table 5.3: Evolution of the two-stage model with $T/\bar{R} = 2/3$ and $\mu(t) = 8$ aircraft per period.

$D(t)$ = number of take-offs during time period t

The quantity $Y(t)$ is the number of aircraft which reach the departure runway during time period t, $Q(t) + Y(t)$ is the total number of aircraft available for take-off during period t, and $\min[Q(t) + Y(t), T\mu(t)]$ is the number of take-offs seen during the period. Finally, $N(t) = X(t) + Q(t)$ is the total number of aircraft rolling out at the beginning of time period t.

EXAMPLE: Table 5.3 and the bottom plot of Figure 5-6 show the evolution of the two-stage model, given the simple push-back schedule with $B(1) = 9$ and $B(2) = 18$. In the first and second periods, the system is not saturated, and all aircraft moving from the taxiways to the departure runway are able to take off immediately. At the start of the third period there are 21 aircraft on the taxiway, and $(2/3)21 = 14$ available for take-off. The system is saturated and only allows eight aircraft to depart, so the remaining six enter the departure queue. Therefore, $X(4) = 7$ and $Q(4) = 6$. During the fourth period the system is again saturated. All six aircraft in the queue take off, but not all aircraft moving from the taxiway to the departure queue can depart. That is, $Y(4) = 14/3$ but only 2 of these depart so $Q(5) = 8/3$.

In terms of the number of aircraft on the airfield, $N(t)$, the evolution of the one-stage and two-stage models is identical until the fifth period, when the two-stage model has 4.2 take-offs while the one-stage model has 3.3. The small difference is due to the immediate availability of all aircraft in the departure queue of the two-stage model. In the one-stage model, these aircraft are assumed to be on the taxiway and only 2/3 of the aircraft are available for take-off. In fact, the difference between $D(5)$ in the two models is exactly one-third of the departure queue.

As in the one-stage model, the *forecast equations* mirror the system equations:

$$\begin{aligned}
\hat{X}(t+1|t) &= \hat{X}(t|t) - \hat{Y}(t|t) + \hat{B}(t|t) && (5.10)\\
\hat{Q}(t|t) &= \hat{Q}(t|t) + \hat{Y}(t|t) - \hat{D}(t|t) \\
\hat{Y}(t|t) &= \frac{T}{R}\hat{X}(t|t) \\
\hat{D}(t|t) &= \min[\hat{Q}(t|t) + \hat{Y}(t|t), T\hat{\mu}(t|t)] && (5.11)
\end{aligned}$$

However, the update equations of the two-stage model are a bit more complex than those of the one-stage model. Given $\hat{X}(t-1|t-1)$ and $\hat{Q}(t-1|t-1)$, and after having observed $B(t-1)$ and $D(t-1)$, we wish to find new estimates $\hat{X}(t|t)$ and $\hat{Q}(t|t)$. It would seem to make sense to add $B(t-1)$ to $\hat{X}(t-1|t-1)$ and subtract $D(t-1)$ from $\hat{Q}(t-1|t-1)$ in order to add push-backs to the count of taxiing aircraft and subtract take-offs from the departure queue. However, the number of observed take-offs is often larger than $Q(t)$, and some proportion of take-offs must also flow directly from the taxiway during the ten minute period.

Our solution is to use the number of observed take-offs in the previous period $D(t-1)$ to find $\hat{Y}(t-1|t)$, an estimate of the number of aircraft that moved from the taxiway to the runway during the previous period. At a minimum, $\hat{Y}(t-1|t) = \frac{T}{R}\hat{X}(t-1|t-1)$, as prescribed by the assumption that taxiing aircraft are spread uniformly along the taxiway. If the observed number of takeoffs $D(t-1)$ is less than the number expected, the remaining aircraft are assumed to remain in the departure queue. If the number of observed take-offs exceeds the total number of aircraft which were expected to have reached the departure queue by the end of the period, then $\hat{Y}(t-1|t)$ is adjusted to include the excess. This motivates the following update equations:

$$\begin{aligned}
\hat{Y}(t-1|t) &= \max[\frac{T}{R}\hat{X}(t-1|t-1), D(t-1) - \hat{Q}(t-1|t-1)] && (5.12)\\
\hat{X}(t|t) &= \hat{X}(t-1|t-1) - \hat{Y}(t-1) + B(t-1) && (5.13)\\
\hat{Q}(t|t) &= \hat{Q}(t-1|t-1) + \hat{Y}(t-1) - D(t-1) \\
\hat{N}(t|t) &= \hat{X}(t|t) + \hat{Q}(t|t)
\end{aligned}$$

where

$\hat{Y}(t-1|t)$ = estimated number of aircraft which flow from the taxiway

t	$N(t)$	$D(t)$	$B(t)$	$\hat{Y}(t-1\|t)$	$\hat{X}(t\|t)$	$\hat{Q}(t\|t)$	$\hat{N}(t\|t)$	$\hat{D}(t\|t)$	$\hat{N}(t+1\|t)$
1	0	0	9	0	0	0	0	0	9
2	9	3	18	0	9	0	9	6	21
3	24	9	0	6	21	3	24	8	16
4	15	9	0	14	7	8	15	8	7
5	6	5	0	4.6	2.3	3.6	6	5.2	0.8
6	1	1	0	1.6	0.8	0.2	1	0.7	0.3
7	0	0	0	0.8	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0

Table 5.4: Updated estimates and one-period forecasts using the two-stage model.

to the departure queue during time period $t-1$

$B(t-1)$ = number of push-backs during time period $t-1$

$D(t-1)$ = number of take-offs during time period $t-1$

The updating equations represent simple 'bookkeeping' to make sure that the sum of the estimated number of aircraft in queue, $\hat{Q}(t|t)$, and aircraft taxiing, $\hat{X}(t|t)$, equals the observed number on the field, $N(t)$.

EXAMPLE: Table 5.4 shows these updates and forecasts for our simple example. The nine push-backs during period 1 immediately enter the taxiway, so $\hat{X}(2|2) = 9$. Of these, the model predicts that 6 will leave the taxiway, pass through the departure queue, and take-off during the second period. However, the actual number of departures during the second period is 3. 3 less than expected, and the remaining three aircraft are placed in the departure queue ($\hat{Q}(3|3) = 3$). During period 2, 18 push-backs occur, so that $\hat{X}(3|3) = \hat{X}(2|2) - 6 + 18 = 21$. These movements from taxiway to departure queue to take-off continues during periods 4 - 6. In each of these periods, $\hat{Y}(t-1|t) = (T/\bar{R})\hat{X}(t-1|t-1)$, since the observed number of departures never exceeds the number of take-offs predicted by the model. However, during period 6 the final aircraft departs, and the update in period 7 sets $\hat{Y}(6|7) = D(6) - \hat{Q}(6|6) = 0.8$ in order to 'pull' the remaining aircraft from the taxiway and clear the airfield.

It is useful to compare the estimates of the one-stage model in the example summarized by Table 5.2 with the two-stage model in Table 5.4. For both models, $\hat{N}(t|t) = N(t)$ for all t. However, the two-stage model separates aircraft into two states, allowing aircraft to

flow from the taxiway into a buffer when the system in saturated. The examples above suggest that the aggregate behaviors of the two models are similar. Large differences only appear under a specific combination of circumstances: when there is a sudden shift from an extremely low to an extremely high capacity and the aircraft flow rate $T/\bar{R}$ is very small, so that the system can make full use of the departure queue buffer of the two-stage model.

5.4 Estimates of Capacity

In the preceding descriptions of the deterministic flow models, an important issue has been ignored. How does one determine the take-off capacity, $\mu(t)$? This section presents four procedures for estimating $\mu(t)$. The first is based on the FAA's capacity standards while the second and third use observed take-off counts to repeatedly adjust the capacity estimates. The final method is motivated by the functional relationship between congestion and take-off rates in Figure 5-5.

(i) *Engineering Performance Standards*

For both the one-stage and two-stage models, let $\hat{\mu}(t|t)$ be the Engineering Performance Standard (EPS) for the runway configuration during period t, the weather category during period t, and a 50%/50% arrival/departure mix. This assumption of a balance between arrivals and departures will frequently be violated in practice. However the EPS numbers do establish a baseline against which to compare the following three capacity estimates.

(ii) *Real-time capacity estimates by exponential smoothing*

In the one-stage model, if the take-off demand, $(T/\bar{R})N(t)$, exceeds the current estimate of capacity during the period, $T\hat{\mu}(t|t)$, then we consider the system to be saturated. When saturation occurs, let the latest observed take-off count $D(t-1)$ be included in a smoothed average which serves as an estimate of the true capacity. In the following update equation for the one-stage model, the smoothing parameter is α, $0 \leq \alpha \leq 1$:

$$\text{if } \left(\frac{T}{\bar{R}}N(t-1) > T\hat{\mu}(t-1|t-1)\right)$$

$$\hat{\mu}(t|t) = \alpha\hat{\mu}(t-1|t-1) + (1-\alpha)\frac{D(t-1)}{T} \tag{5.14}$$

else

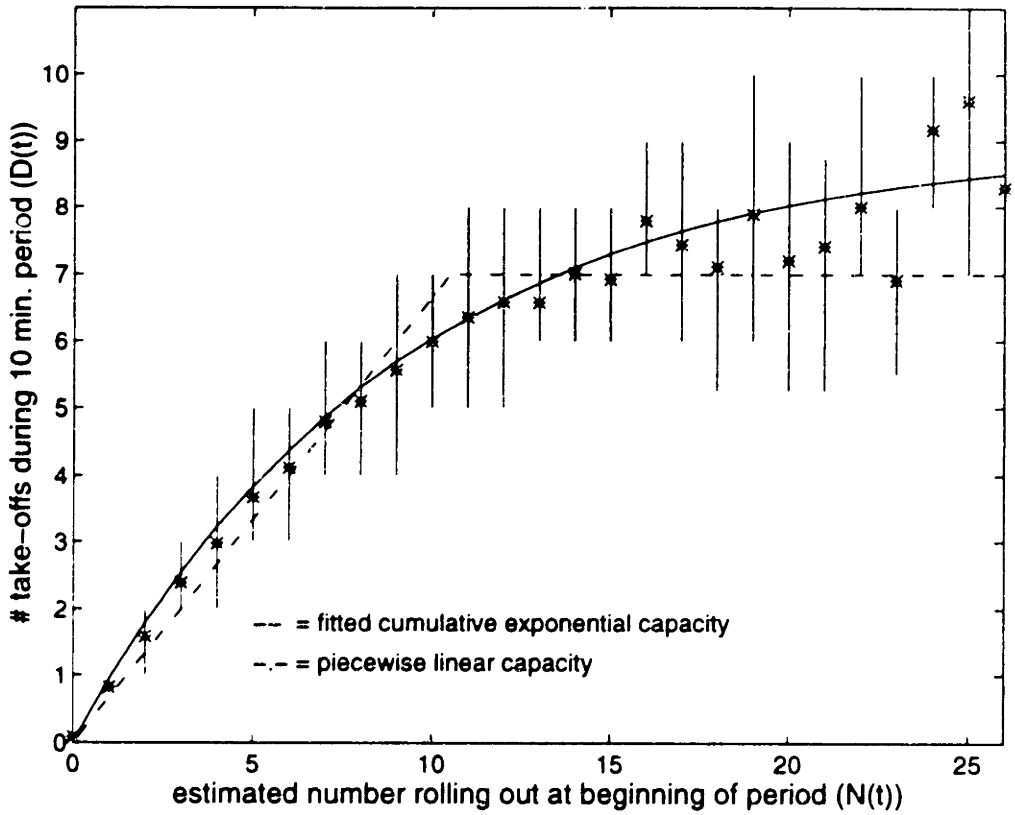


Figure 5-8: Estimated number rolling out (N(t)) and observed number of take-offs during each period ($D(t)$), along with fitted piecewise linear and cumulative exponential capacity estimates.

$$\hat{\mu}(t|t) = \hat{\mu}(t-1|t-1) \tag{5.15}$$

The update equation for the two-stage model is almost identical, with a small change in the 'if' statement that tests for saturation. Replace the quantity $(T/\bar{R})N(t-1)$ with $(T/\bar{R})X(t) + Q(t)$.

With this capacity function, we are assuming that the functional relationship between $N(t)$ and $D(t)$ is piecewise linear, with a plateau at $T\hat{\mu}(t|t)$. The relationship is shown by the dotted line in Figure 5-8 with $T\hat{\mu}(t|t) = 7$. As can be seen in the figure, the 'capacity' is often exceeded, which is not surprising, given Newell's point that observed rates may fluctuate around a long-term, sustainable capacity. The purpose of the exponential smoother is to separate temporary fluctuations from longer-term changes in capacity.

(iii) *Real-time capacity estimates by Kalman Filtering*

Another method for smoothing over recent take-off rates is by iteratively fitting a *dy-*

namic linear model (DLM) for $\mu(t)$ with the real-time data. While the DLM itself is an unlikely description of the evolution of $\mu(t)$ capacity over time, the model provides a framework for deriving capacity updating equations with many useful properties. The DLM describes the evolution of $\mu(t)$ as a stochastic process, where the current capacity level is a linear function of the level in the previous period. Uncertainty about the current level of $\mu(t)$ is introduced by adding a random shock:

$$\mu(t+1) = \mu(t) + \omega(t) \tag{5.16}$$

When the system is saturated, the observed take-of count is proportional to the true capacity, plus some added noise:

$$D(t) = T\mu(t) + \nu(t) \tag{5.17}$$

In the simplest case, both $\omega(t)$ and $\nu(t)$ are uncorrelated zero-mean white noise processes with constant variances W and V, respectively.

The minin m mean squared error estimates of $\mu(t)$, given take-off counts ($D(1)$, $D(2)$, ..., $D(t)$), can be calculated from a well-known set of equations, the Kalman Filter. When the system is saturated, the filter updating equations are:

$$\text{if} \quad \left(\frac{T}{R}N(t) > T\hat{\mu}(t|t)\right)$$

$$\hat{\mu}(t|t) = \hat{\mu}(t-1|t-1) + G(t)\left[D(t) - T\hat{\mu}(t-1|t-1)\right] \tag{5.18}$$

$$P(t) = P(t-1) - G(t)^2 Q(t) \tag{5.19}$$

where

$$Q(t) = T^2(P(t-1) + W) + V$$

$$G(t) = T(P(t-1) + W)/Q(t).$$

The quantity $P(t)$ is the variance of our latest estimate of $\mu(t)$. The term $G(t)$ is often referred to as the *Kalman gain*. This term governs the rate of adaption of the estimate to new data. As $G(t)$ grows, recent observations are given more weight. The quantity $1 - \alpha$ plays a similar role in the exponential smoother.

When the system is not saturated, we do not obtain any information that would lead

us to revise our capacity estimates, and $\hat{\mu}(t|t)$ is not changed:

$$\begin{aligned} \text{if} \qquad & \left(\frac{T}{R}N(t) \leq T\hat{\mu}(t|t)\right) \\ \hat{\mu}(t|t) &= \hat{\mu}(t-1|t-1) \qquad (5.20) \\ P(t) &= P(t-1)+W. \qquad (5.21) \end{aligned}$$

If the system were always saturated, $G(t)$ approaches a constant as t increases. The limiting form of the Kalman Filter is equivalent to equation 5.14 with $1-\alpha$ replaced by the constant gain G, which is a function of the variances of the white noise processes ([41], p. 89):

$$\begin{aligned} G &= \lim_{t\to\infty} G(t) \\ &= \frac{TW}{2V}\left[\sqrt{1+(4V)/(WT^2)}-1\right] \qquad (5.22) \end{aligned}$$

Note that as the observation variance V grows larger, the gain grows smaller and recent observations are given less weight. As the system variance W grows larger, the gain and the rate of adaption to new data grow as well. With W large, capacity varies widely over time and we would want the most recent observations to carry more weight.

In our application, however, there are periods of non-saturation, and during these periods both the variance $P(t)$ and the gain $G(t)$ grow. This natural 'softening' of the estimate over unsaturated periods is one potential advantage of the filter over the exponential smoother. In addition, the filter allows us to distinguish between take-off variation due to random fluctuations (represented in the equations by $\nu(t)$ with variance V) and capacity variation due to more gradual changes in capacity (represented by $\omega(t)$ with variance W). The Kalman Filter may also be adapted to a more general model, in which the perturbations $\omega(t)$ and $\nu(t)$ are serially correlated or correlated with each other. This would reflect the behavior of airport capacity, which moves up and down over time according to changing weather conditions.

(iv) *Cumulative Exponential Capacity Estimates*

Despite its relative complexity, the Kalman filter assumes the same piecewise linear relationship between demand and capacity as the EPS and exponentially smoothed esti-

mates. The empirical relationship between $N(t)$ and $D(t)$ exhibited in Figure 5-8 suggests a smoother relationship than the piecewise linear function allows. There is no clear transition between free-flow and saturation, and $D(t)$ is highest at the highest levels of $N(t)$. A function which traces a curve similar to that of the figure is:

$$d(N(t)) = \mu_{max}\left(1 - e^{-\lambda N(t)}\right) \tag{5.23}$$

where μ_{max} and λ are parameters of the function. The function has an asymptote at μ_{max} and the rate of increase is determined by λ. The function is a scaled cumulative distribution function for an exponential random variable, and we will refer to this as the *cumulative exponential capacity estimate.*

We performed a grid search to find the parameter values for $d(N(t))$ that minimize the sum of squared residuals between $d(N(t))$ and the observed take-offs. We found the values of μ_{max} and λ that solved the following problem:

$$\min_{\mu_{max},\lambda} \sum_{t} [D(t) - d(N(t))]^2 \tag{5.24}$$

The optimal parameter values are $\mu_{max} = 9.0$ and $\lambda = 9.2$ and the curve generated by these values is displayed in Figure 5-8.

To this static model of the relationship between demand and departure rates we add a dynamic term, $\hat{K}(t|t)$, which adjusts the curve up or down, depending on the most recent observations of the take-off rate. Given a value of $N(t)$, our expected take-off count is:

$$\hat{D}(t|t) = d(N(t)) + \hat{K}(t|t) \tag{5.25}$$

The adjustment $\hat{K}(t|t)$ is updated by smoothing over the residuals of the cumulative exponential capacity estimate:

$$\hat{K}(t|t) = \hat{K}(t-1|t-1) + \alpha'[D(t-1) - d(N(t-1))] \tag{5.26}$$

The cumulative exponential capacity estimate is a significant change from the one-stage and two-stage models. We have dispensed with the piecewise linear departure predictions, which were justified by the assumption that aircraft are spread uniformly along the taxi-way.

Instead, we assume a relationship between demand and take-off rates which is described by the cumulative exponential function but is not justified by any physical model of the departure system. In fact, for certain values of $\hat{K}(t|t)$, it is possible to predict a $\hat{D}(t|t)$ which is less than zero. However, the function provides a good fit to the data. In the next section we will investigate how well these capacity estimation procedures predict actual departure rates and airfield congestion.

5.5 Empirical Tests of Departure Flow Models

The models described in the previous sections were implemented to test their forecast accuracy. Forecasts of $N(t)$, the number of aircraft rolling out at time t, were generated from both the one-stage and two-stage models using each of the four capacity estimation procedures. Data to test the models were generated from the sequences $B(t)$ and $D(t)$. These sequences were calculated by counting the number of push-backs and take-offs during 3,105 ten-minute periods in March and August, 1991. Since push-back times were not known for all flights, the $B(t)$ and $N(t)$ were estimated using the procedure described in Section 5.2. The values of $N(t)$ are not the actual number rolling out, for they are calculated from inaccurate information about aircraft push-backs and take-offs. When calculating forecast errors, we will be comparing the *estimate* $N(t)$ with a forecast of that estimate. However, in Section 5.2 we saw that these values of $N(t)$ are a rough lower bound on the actual number rolling out, while a rough upper bound was, on average, within 0.5 aircraft of this lower bound. Therefore, we are confident that $N(t)$ is close to the actual number rolling out. Even when the values of $N(t)$ are not precise, these experiments should prove useful for comparing the relative merits of the forecasting procedures.

As an example, Figure 5-4 displays the levels of $N(t)$ and $D(t)$ during the morning of August 27, 1991. Figure 5-9 displays the values of $N(t)$ over the entire day. The Figure also shows the 30-minute forecasts of $N(t)$ produced by the one-stage flow model using exponentially smoothed capacity estimates. Each forecast, $\hat{N}(t+3|t)$, was produced by recursively evaluating the forecast equation 5.7, which simply adds expected push-backs to $N(t)$ and subtracts expected take-offs. The forecasts remain fairly close to the observed values in the morning, but deviate by as much as eight aircraft in the afternoon. When averaged over all 27 days (3,105 ten-minute periods), the forecast residuals $N(t+3) - \hat{N}(t+$

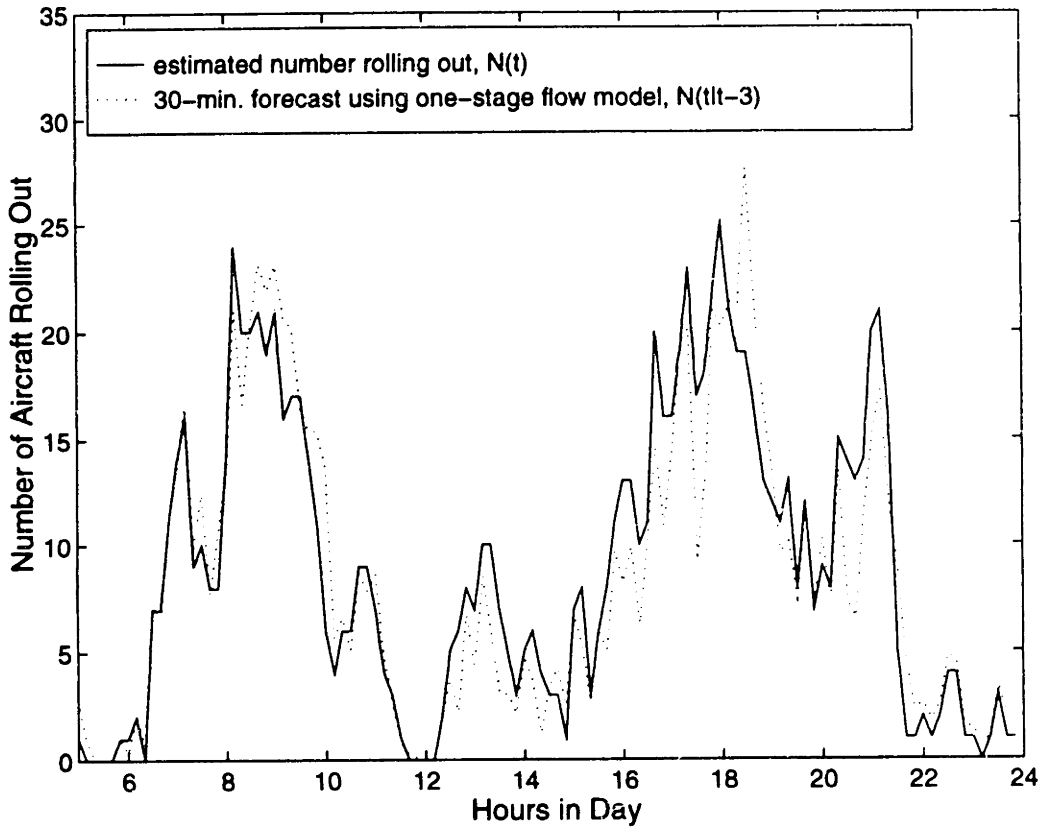


Figure 5-9: Estimated number of aircraft rolling out ($N(t)$) and predicted number rolling out at Logan Airport during August 27, 1991. Predictions were generated by the one-stage flow model using exponential smoothing to estimate capacity.

$3|t$) can be used to estimate the forecast accuracy of the model.

While Figure 5-9 only shows 30-minute forecasts, we calculated forecasts over horizons ranging from $k = 1$ (ten-minute forecasts) to $k = 6$ (one-hour forecasts). Forecast accuracy for each model over each forecast horizon will be summarized with four statistics. The first two are the mean forecast error and root-mean squared forecast error (RMSE):

$$\begin{aligned} \text{mean forecast error} &= \frac{\sum_{t=1}^{N-k}[N(t+k) - \hat{N}(t+k|t)]}{N-k} \\ \text{RMSE} &= \sqrt{\frac{\sum_{t=1}^{N-k}[N(t+k) - \hat{N}(t+k|t)]^2}{N-k}} \end{aligned} \tag{5.27}$$

However, these averages may not accurately reflect the effectiveness of the algorithms. The simple mean gives the same weight to forecasts for time periods with little or no departure activity (i.e. 5:30 a.m.) as it does to forecasts for peak departure periods. A

more representative statistic is the mean error seen by aircraft on push-back. A close approximation to this is a weighted mean forecast error in which the weight assigned to each error is the number of push-backs during the surrounding ten-minute period:

$$\begin{aligned}&\text{weighted mean forecast error}\\ &= \frac{\sum_{t=1}^{N-k} B(t+k)[N(t+k) - \hat{N}(t+k|t)]}{\sum_{t=1}^{N-k} B(t+k)} \qquad (5.28)\\ &\text{weighted RMSE}\\ &= \sqrt{\frac{\sum_{t=1}^{N-k} B(t+k)[N(t+k) - \hat{N}(t+k|t)]^2}{\sum_{t=1}^{N-k} B(t+k)}} \qquad (5.29)\end{aligned}$$

These statistics will be used to compare the one-stage and two-stage models, as well as the four methods for estimating take-off capacity.

Two more important modeling simplifications should be mentioned before the results of empirical tests are presented. First, we use the same push-back times to calculate the 'real' $N(t)$ and to generate forecasts, so that $\hat{B}(t+k|t) = B(t+k)$ for all t and k. This implies that forecast errors for $N(k)$ are due solely to errors in predictions of take-off rates over the periods in the forecast horizon:

$$N(t+k) - \hat{N}(t+k|t) = \sum_{l=0}^{N-k} \left[\hat{D}(t+l|t) - D(t+l)\right] \qquad (5.30)$$

In other words, since there are no errors in push-back forecasts, the k-step forecast error of $N(t)$ is an accumulation of errors in individual take-off rate forecasts over k periods. The assumption of perfect push-back forecasts may exaggerate the accuracy of these models. This assumption will be relaxed, and its effects explored, in Section 5.5.3.

The second simplification concerns the value of the model parameter, $\bar{R}$, the taxi-time of each aircraft. For both the one-stage and two-stage models, the value of the $\bar{R}$ is set to a constant 15 min. Of course, one wonders whether forecast accuracy is sensitive to the value of $\bar{R}$, and the effect of changes to $\bar{R}$ will also be tested in section 5.5.3.

5.5.1 Comparison Between One-Stage and Two-Stage Models

Empirical tests with the Logan Airport data indicate that the one-stage and two-stage models produce almost identical predictions of $N(t)$. For example, using EPS capacity estimates, the mean absolute deviation in one-hour forecasts between the two models averages

Error Statistic	Capacity Estimate	forecast horizon (min.) 10	20	30	40	50	60
Mean	EPS	0.0	0.1	0.1	0.1	0.1	0.1
	SMOOTH	0.0	0.0	0.0	0.0	0.0	0.0
	KALMAN	0.0	0.0	0.0	0.0	0.0	0.0
	CUMEXP	0.0	0.0	0.0	0.0	0.0	0.1
RMSE	EPS	1.6	2.1	2.5	2.9	3.2	3.4
	SMOOTH	1.4	1.8	2.1	2.3	2.4	2.6
	KALMAN	1.4	1.8	2.0	2.2	2.4	2.5
	CUMEXP	1.4	1.7	1.9	2.0	2.1	2.2
Weighted Mean	EPS	0.2	0.3	0.3	0.4	0.3	0.3
	SMOOTH	0.0	0.0	0.1	0.1	0.1	0.1
	KALMAN	0.0	0.0	0.1	0.1	0.1	0.1
	CUMEXP	0.0	0.1	0.1	0.1	0.2	0.2
Weighted RMSE	EPS	1.8	2.5	3.1	3.5	3.9	4.3
	SMOOTH	1.6	2.1	2.4	2.7	2.9	3.1
	KALMAN	1.6	2.0	2.4	2.6	2.8	3.0
	CUMEXP	1.5	1.9	2.1	2.3	2.4	2.5

Table 5.5: Forecast error statistics of the one-stage model for predictions of $N(t)$, evaluated with EPS capacity estimates, exponentially smoothing (SMOOTH), Kalman Filtering (KALMAN) and the cumulative exponential capacity estimates (CUMEXP). Error statistics for the two-stage model were virtually identical.

0.05 aircraft, and has a maximum value of 2.4 aircraft. The mean one-hour forecast error for both models is 0.1 min., and the RMSE for one-hour forecasts of both models is 3.4 min. The rows labeled 'EPS' in Table 5.5 display the mean and weighted mean errors for the one-stage model over six forecast horizons using EPS capacity estimates. These aggregate statistics, as well as those produced by the other capacity estimation procedures, were all within 0.1 min. of the error statistics collected for the two-stage model.

Based on the available information, it is impossible to say whether the one-stage or two-stage model is a more accurate description of the dynamics of the departure process. Information about the distribution of aircraft along the taxiways would distinguish them. The one-stage model assumes that aircraft rolling out are distributed uniformly along the taxiway. In the two-stage model, aircraft which find a saturated system are held in a departure queue and may be served as soon as the departure runway is available. Observations of Logan Airport indicate that the truth is somewhere between these extremes. When demand for departure runways exceeds capacity, aircraft do tend to collect in a queue near

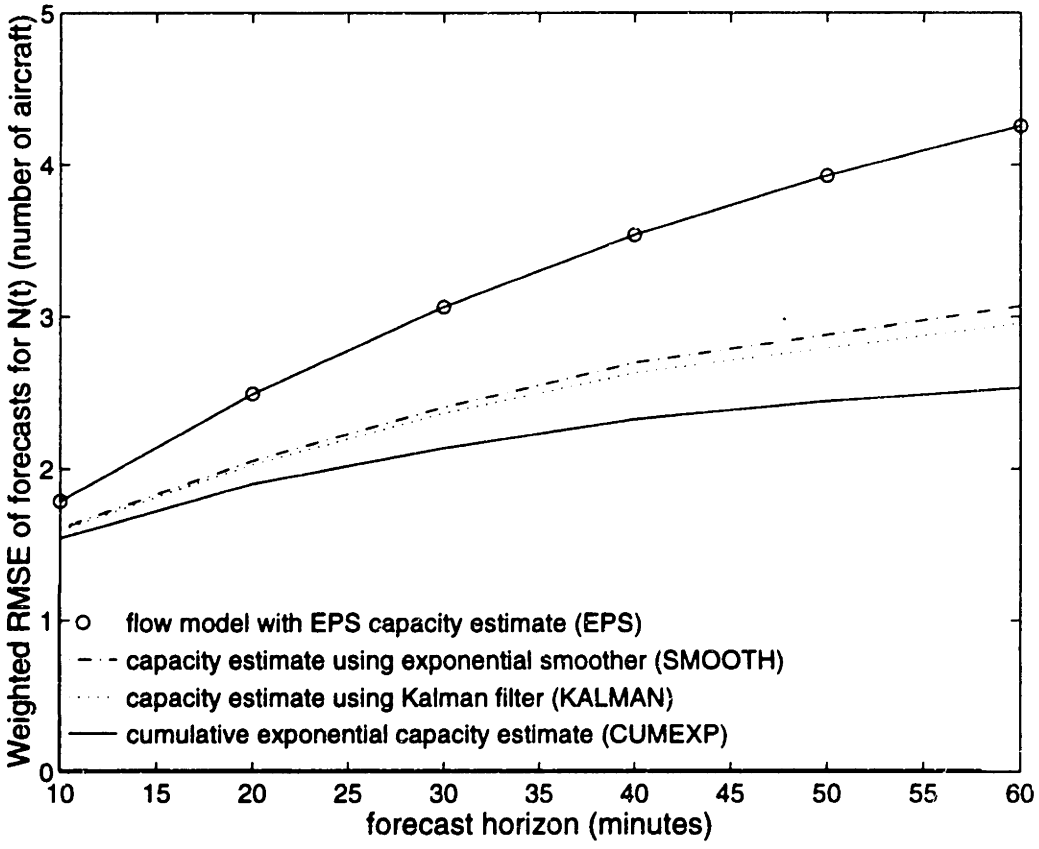


Figure 5-10: Root mean squared forecast errors (RMSE) for four methods of estimating take-off capacity.

the departure runway. However, at these times many aircraft may be delayed on the ramps near the gates or at intersections between taxiways and runways. Given a departure queue at push-back, pilots may see no reason to 'hurry up and wait' and therefore may taxi more slowly. Since we are interested in predicting $N(t)$, and the two models have almost identical predictive power, we will adopt the simpler one-stage model for all subsequent testing.

5.5.2 Comparisons Among Capacity Estimation Procedures

Using the one-stage model, we compare our four methods for estimating take-off capacity: the Engineering Performance Standards (EPS), the real-time estimates from the exponential smoother (SMOOTH), the estimates from the Kalman Filter (KALMAN) and the estimates produced by the cumulative exponential capacity estimate (CUMEXP).

Each of the last three capacity estimation procedures require the setting of parameters which govern the sensitivity of the estimate to the most recent observations. A numerical search found that 0.9 is the optimal value of α in the exponential smoother. That is, after observing $D(t)$ in a saturated system, the next capacity estimate gives 10% of its weight to this observation and 90% to the previous capacity estimate. Optimal values of the Kalman Filter parameters were also determined by numerical search. The values used were $\text{var}(\omega(t)) = W = 0.05$ and $\text{var}(\nu(t)) = V = 3$ for all t. This implies that the variance of the actual take-off counts around the 'true' capacity in the model is 3 aircraft per ten-minute period (a standard deviation of 1.73 aircraft). In our implementation of the filter, $\mu(t)$ was expressed as an hourly rate, so the variance $W = 0.05$ implies that capacity has a standard deviation of $\sqrt{6(0.05)} = 0.55$ aircraft/hour from hour to hour. This may seem small, but the quantity which determines the adaptability of the filter is the ratio W/V. In the limit, the value of the Kalman gain G calculated from equation 5.22 is 0.13, which is comparable to the value of $1 - \alpha = 0.1$ in the exponential smoother.

Three parameters govern the cumulative exponential capacity estimate. As described in section 5.4, μ_{max} and λ were set to 9.0 and 9.2, respectively. The smoothing parameter in equation 5.25 was set to $\alpha' = 0.9$, as in the exponential smoother.

Table 5.5 and Figure 5-10 display the mean forecast errors of these four methods for forecast horizons from 10 to 60 minutes. The mean forecast errors of the real-time capacity estimates are close to zero, indicating that the models produce unbiased forecasts of $N(t)$. However, the EPS-based forecasts tend to underestimate $N(t)$ and therefore are overes-

timating the take-off rate. The weighted root mean squared errors shown in figure 5-10 indicate that the forecasts based on real-time capacity estimates are significantly more accurate than those based on the EPS numbers, particularly for the long forecast horizons. With a forecast horizon of one hour, both the exponential smoother and the Kalman filter improve the root mean squared error by over one aircraft. The cumulative exponential capacity estimate improves one-hour forecasts by an additional 0.5 aircraft.

Figure 5-11 displays both the EPS and exponentially smoothed capacity estimates for four days in August, 1991. In general, the real-time estimates are more stable than the EPS values, which vary widely over time. However, without more information it is unwise to generalize from these pictures. The EPS values used here are estimates of capacity which are based on our crude measurements of weather conditions: hourly ceiling and visibility measurements, classified into one of four weather states. In addition, we assume a 50/50 mix of departures and arrivals, an assumption which is frequently violated in practice. These empirical tests do highlight a significant advantage of the real-time capacity estimates: they do not depend on accurate weather observations, information about runway configurations, or knowledge of the arrival/departure mix.

Now consider the RMSE of the forecasting models as the forecast horizon grows. The CUMEXP model provides the most accurate forecasts, with a RMSE of 1.4 and 2.2 aircraft over 10-minute and 60-minute forecast horizons, respectively. At first glance, this rate of growth in forecast error is puzzlingly slow. Each k-step forecast error is the sum of k take-off forecast errors (see equation 5.30). If each take-off forecast error is independent and has the same distribution, then the standard deviation of $N(t+k|t)$ should rise with the square root of k. Let σ_D be the standard deviation of a single take-off rate forecast $\hat{D}(t+k|t)$. From equation 5.30, the RMSE of a 10-minute forecast $\hat{N}(t+1|t)$ should suffice as an estimate of σ_D. Therefore, we might expect:

$$\begin{aligned} \text{Std. Dev.}\left[(\hat{N}(t+6|t)\right] &= \sqrt{6}\sigma_D \\ &\approx \sqrt{6}\ \text{Std. Dev.}\left[(\hat{N}(t+1|t)\right] \end{aligned} \tag{5.31}$$

From Table 5.5, for example, we would expect the six-period RMSE for the cumulative exponential algorithm to be $\sqrt{6}(1.4) = 3.4$ aircraft. The actual six-period RMSE, however, is only 2.2 aircraft.

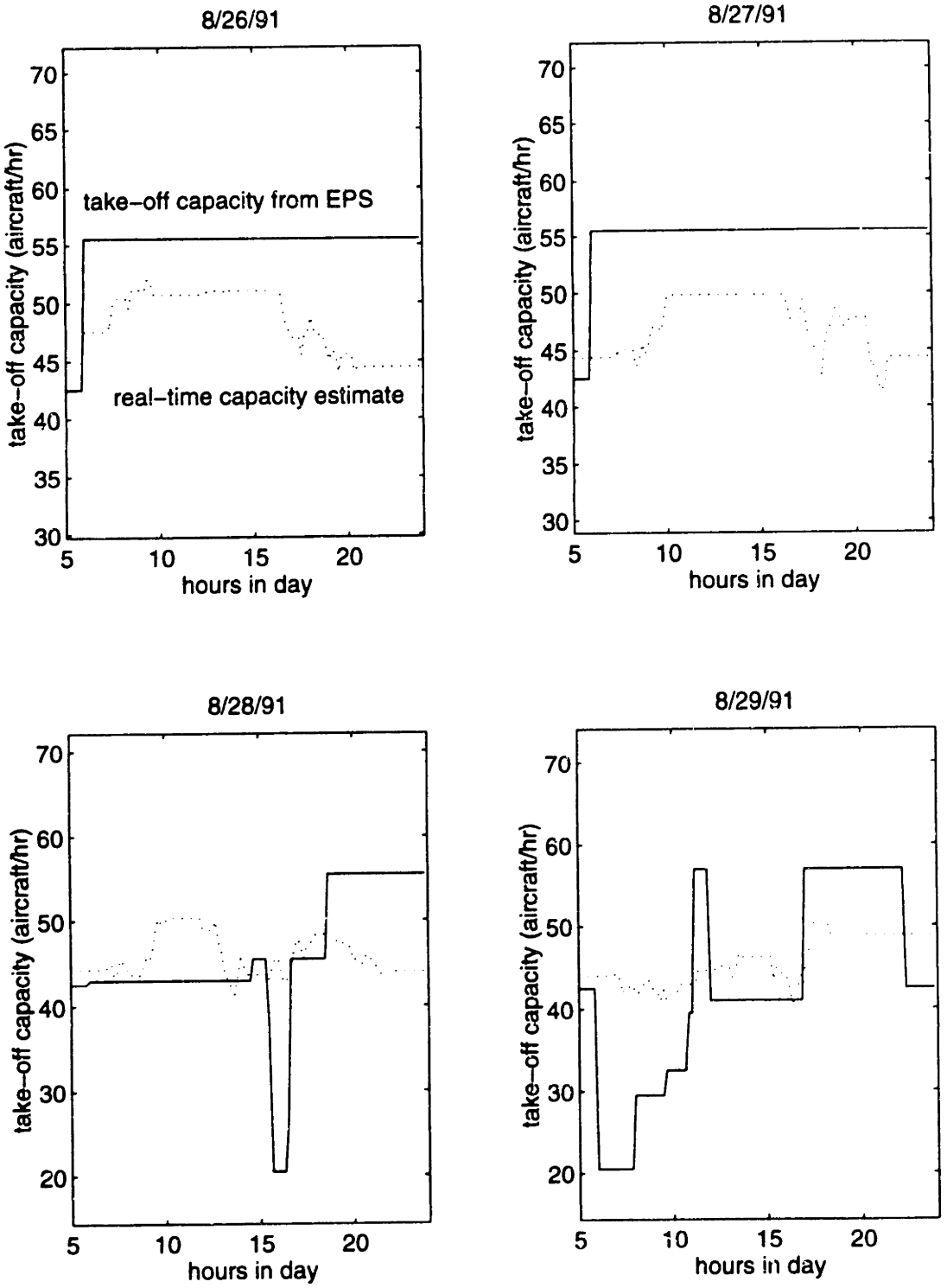


Figure 5-11: EPS numbers and real-time take-off capacity estimates at Logan Airport for four days in August, 1991

The slow growth of forecast error is due to the fact that take-off errors are *not* independent, but are inversely correlated. By comparing the k-step departure rate forecast with the $k+1$st, we find the correlation coefficient:

$$\rho(\hat{D}(t+k|t), \hat{D}(t+k+1|t)) \approx -0.2 \tag{5.32}$$

This has an intuitive interpretation: an optimistic take-off prediction in one period will often be followed by an underestimate of the take-off rate in the next. As long as the total number of aircraft which flow through the model is correct, there is a limited number of aircraft to send into the air at the wrong time.

Finally, the model residuals were scrutinized to determine if there are any apparent deficiencies in the model. For all three models with real-time capacity estimates (SMOOTH, KALMAN, and CUMEXP), there was a slight autocorrelation in the model errors from time period to time period:

$$\rho(\hat{N}(t+1|t), \hat{N}(t+2|t+1)) \approx 0.04 \tag{5.33}$$

The autocorrelation implies that there may be some time-varying aspect of the departure process which is not captured by the model. However, the autocorrelation is quite small and did not justify further changes in the model form.

5.5.3 Sensitivity to Modeling Simplifications

The forecasts described above were generated using 'perfect' gate departure delay information, in the sense that the same push-back times, PB2, were used to both calculate and predict $N(t)$.[2] This is a 'best case,' and true push-back forecasts are likely to be somewhat inaccurate. We saw in Chapter 4 that the accuracy of gate departure delay forecasts may vary widely, and depends on the information that is available to the forecaster. In order to measure the sensitivity of the model to this best-case assumption we also derived 'worst-case' push-back forecasts, PB3(i), based only on the scheduled departure times and

[2]Of course, the push-back time estimates PB2 themselves are not perfect, but were estimated from the available information according to the procedure described in section 5.2.1. However, knowledge of PB2 provides perfect information about the demand pattern used to generate $N(t)$, and it is $N(t)$ which is being used to test the models.

the overall average gate departure delay for matched flights:

$$\text{PB3(i)} \quad = \quad \text{SCH(i)} + 5.3 \text{ min.} \tag{5.34}$$

With push-back forecast PB3, we are assuming that no information is available about any push-back other than its scheduled departure time and an overall average delay. The simple turn models described in Chapter 4, for example, should produce substantially more accurate forecasts of push-back times than PB3.

Push-back predictions PB3(i) were calculated for all flights, and forecasts of push-back counts $\hat{B}(t+k|t)$ were derived from these push-back times. We then used these demand forecasts as input to the one-stage model with exponentially smoothed capacity forecasts. Not surprisingly, forecasts of $N(t)$ based on PB3(i) were inferior to those based on perfect push-back forecasts. Over a ten-minute forecast horizon, the weighted RMSE doubles when PB3 are used (1.6 aircraft with 'perfect' push-back forecasts, 3.2 aircraft with PB3). The effect of poor push-back (or gate departure delay) forecasts diminishes as the time horizon grows (see Figure 5-12). With a forecast horizon of one hour, forecasts based on PB3 have a RMSE that is only 30% larger than those based on perfect push-back forecasts. Predictions made with PB3 exhibit strongly the characteristic noted in the previous section: the initial forecast error and subsequent errors are inversely correlated so that initial errors are sometimes corrected in forecasts over longer horizons by errors in the other direction. Forecast errors grow slowly over time.

Finally, we tested the sensitivity of the models to the value of the constant taxi-time, $\bar{R}$. In all previous tests, $\bar{R}$ was set to 15 min., so that 2/3 of the aircraft on the taxiway were available for take-off during each ten-minute period. Forecasts were produced by the one-stage, exponential smoothing model with values of $\bar{R}$ ranging from 10 min. to 25 min. Figure 5-13 displays the forecast accuracy for 10 and 60-minute forecasts. The figure demonstrates that for short forecast horizons the model is insensitive to this parameter, with mean and root mean squared errors varying very little as $\bar{R}$ varies. For the one-hour forecast horizon, reducing the value of $\bar{R}$ by five minutes, or increasing it by ten minutes, only increased the RMSE by one aircraft. However, increasing $\bar{R}$ by ten minutes introduced a significant bias into the model. As $\bar{R}$ is increased, the modeled flow of aircraft from the taxiway to the runways is slowed, and the forecasts of $N(t)$ become artificially high.

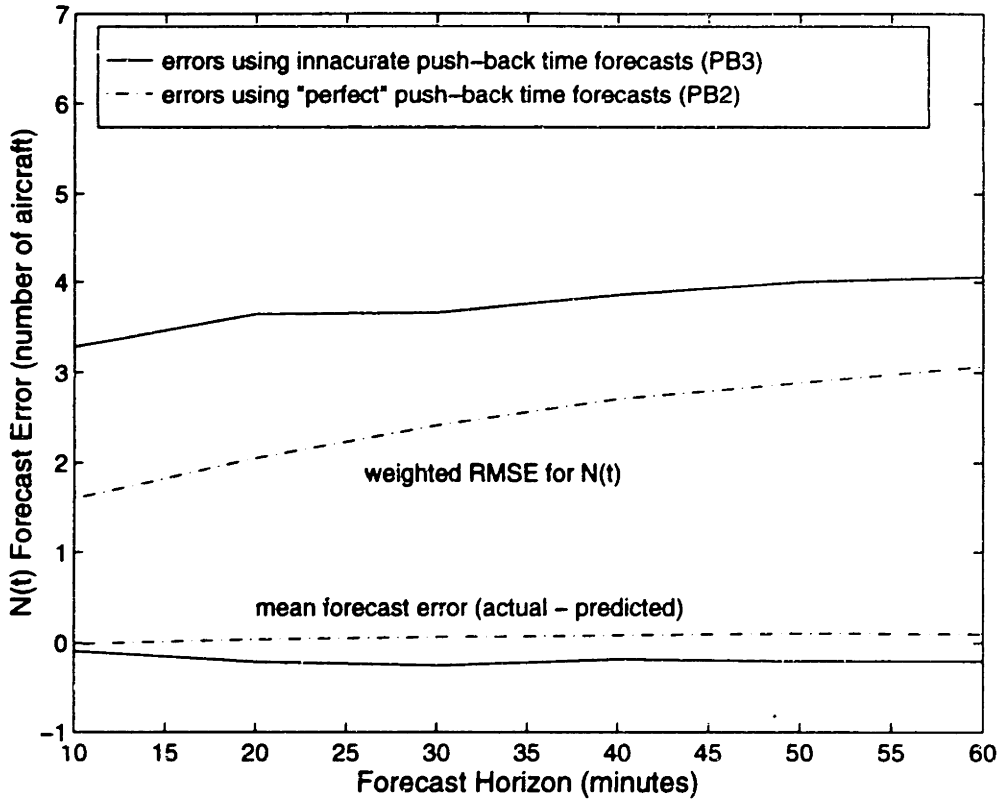


Figure 5-12: Mean forecast errors and root mean squared errors for flow models using two forecasts of push-back times: adjusted scheduled departure times (PB3) and perfect information.

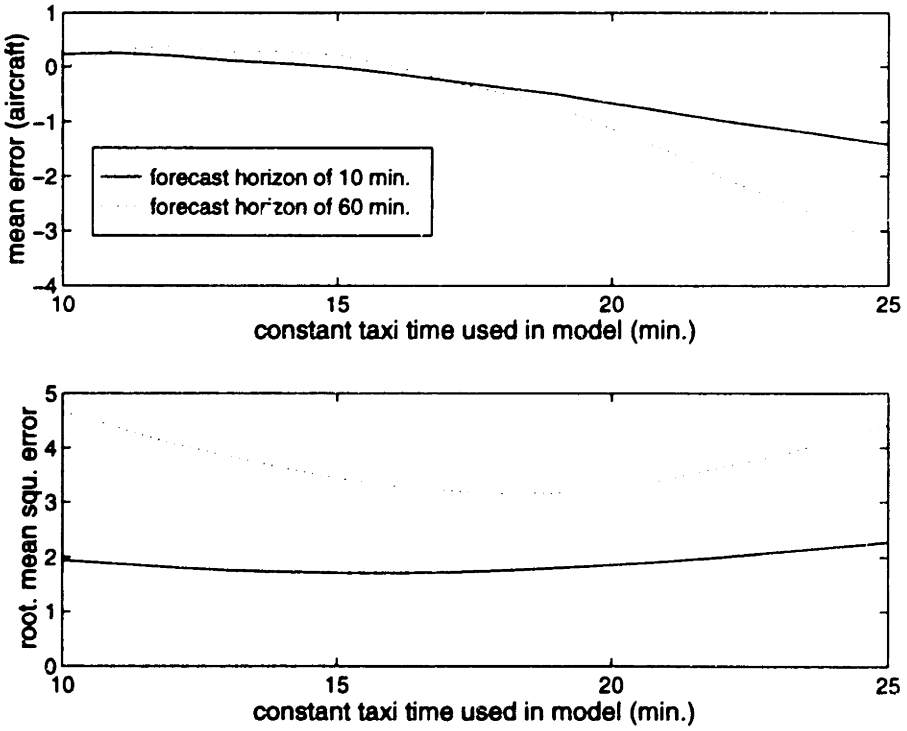


Figure 5-13: Sensitivity of one-stage flow model to variations in constant taxi-time $\bar{R}$

5.6 Conclusions and Extensions

The aggregate flow models offer a simple and accurate method for predicting the number of aircraft rolling out over a short time horizon. The models approximate all departure runways as a single, aggregate runway, and assume that aircraft 'flow' to take-off according to a deterministic process. These assumptions lead to a model which can be easily updated as real-time data arrive. The empirical experiments described in the previous section quantify the forecast accuracy of these models. Given perfectly accurate predictions of push-back times at Logan, it is possible to predict the number of aircraft on the airfield ten minutes and one hour in advance with RMSE of 1.4 and 2.2 aircraft, respectively.

Some proportion of these forecast errors may be caused by random fluctuations in take-off rates around the true capacity of the airport, and some may be due to errors in the data used to test the models. The take-off counts $D(t)$ were calculated from take-off messages which are sometimes inaccurate. The push-back counts $B(t)$ were estimated from unreliable data about push-backs of general aviation and commuter aircraft.

The remaining forecast errors, however, may be diminished by improvements to the models. Three areas for improvement are listed below:

1. **Improved Models of Queueing and Capacity:**

   The cumulative exponential capacity estimate seems to closely approximate the relationship between congestion and take-off rates, but is not justified by any physical model. Therefore, we have no assurance that the relationship is valid for airports other than Logan, or for Logan under very different conditions. An important area of research is the formulation of more detailed statistical models of queueing and capacity that are based on the 'physics' of airport operations. Methods for real-time updating of these models may be derived. It may be useful to incorporate the following effects into the models:

   - interactions between arriving and departing aircraft;
   - particular runway configurations, such as those with two independent, active runways;
   - queues of aircraft that form when taxiing aircraft cross active runways.

   However, given the data currently available to the ETMS, it would be difficult to either

formulate or update these detailed models. In fact, the data from the ETMS was not complete enough to update the simple one-stage or two-stage models (estimates were used when the data were not available). More detailed data, such as those provided by ground surveillance systems, will make more accurate models possible.

2. **Dynamic Ground Control Models:**

   The one-stage and two-stage models assume that taxiing aircraft arrive to the runway at a rate that is proportional to the number of aircraft taxiing. This is justified under the assumption that taxiing aircraft are distributed uniformly along the taxiway. However, more detailed models of the airfield may allow us to improve this model, perhaps by taking into account the dynamic effects of airport ground control. For example, the arrival rate to a departure queue may be influenced by the length of the queue itself, since ground controllers may hold aircraft at the gate or route aircraft to less congested runways when congestion builds.

3. **Modeling the Stochastic Queueing Process:**

   One source of model error is the assumption of a deterministic flow of aircraft to the departure runway. This source of error is explored by Koopman [24], who compares a time-varying queueing system with a deterministic demand process with a similar system driven by Poisson process. He finds that the expected queue lengths derived from the deterministic system can be an inaccurate approximation of the expected queue length of the stochastic system. It may improve the accuracy of the model to add a stochastic component to the description of the departure process in order to take random changes in taxi-times and taxiway queues into account.

   However, the system described in our model does differ in important ways from Koopman's. It would be unreasonable to describe aircraft reaching the departure runway with a Poisson process when we know exactly how many aircraft have pushed back from the gate and are headed for the departure runway. In addition, when an airport is saturated and a departure queue has developed, the deterministic aspects of the system tend to dominate ([26], p. 21). In fact, our deterministic models showed no forecast bias, while the deterministic models described in Koopman's paper dramatically underestimate queue lengths.

Besides generating take-off time forecasts, these statistical models also provide insight

into the overall performance of the airport. The statistical models distill the real-time information flowing from the airport into a few summary parameters. By monitoring these parameters, analysts may gather information about the effects of weather, runway conditions, and other factors on airport performance. They may also compare the operational 'reality,' as described by the statistical model, with the performance predicted by the more standard, static models of airport capacity.

Finally, Logan airport operations are characterized by complex interactions between arrival and departure runways. Model formulation and performance may be significantly different at airports where arrival and departure operations are independent, such as Atlanta Hartsfield and Denver International Airport. The development of models for Atlanta may be especially timely, given the current work on the *Atlanta Airport Resource Management Tool* by the Atlanta Tower and the carriers [2]. At this stage, the project at Atlanta has focused on communication between the airport tower and the ramps as well as the display of data to tower and ramp controllers. However, the methods developed in this chapter have the potential to make a significant contribution to this effort as well as to the ETMS.

Chapter 6

Roll-out Time Models

The models of the previous chapter assigned a common taxi time, R, to all flights and ignored distinctions among aircraft. This simplified the aggregate flow models, but is antithetical to our eventual goal: the production of take-off time forecasts for individual flights. For example, when runway 9 is the primary departure runway, USAir flights at Logan taxi a shorter distance from gate to runway than Northwest flights, and our model should reflect this difference (see Figure 6-1). The models introduced in this chapter produce roll-out time forecasts that take these individual differences into account.

Assume that the flights in our sample are ordered by take-off time. In general, the model for the roll-out time of the kth flight, r_k, has the linear form:

$$r_k = \mathbf{F}_k' \boldsymbol{\Theta}_k + \epsilon_k \tag{6.1}$$

where $\mathbf{F}_k$ is a column vector of explanatory variables (or 'factors'), $\boldsymbol{\Theta}_k$ is a column vector of model parameters, and ϵ_k is the residual roll-out time not explained by the known factors. This chapter describes both the selection of the model factors $\mathbf{F}_k$ and the estimation of the parameters $\boldsymbol{\Theta}_k$. Two types of models will be introduced: *dynamic* and *static*. In equation 6.1, the coefficients $\boldsymbol{\Theta}_k$ are indexed by k, implying that they may vary from one flight to another. This is a dynamic linear model in which the effects of a particular factor, such as bad weather, may vary from flight to flight. An alternative is a static linear model, in which the coefficients $\boldsymbol{\Theta}_k$ are replaced by the vector $\boldsymbol{\Theta}$ which is assumed to be constant over all flights and all time. Forecasts from the 'static' model do change from flight to flight, according to the conditions under which the flight operates. However, the static

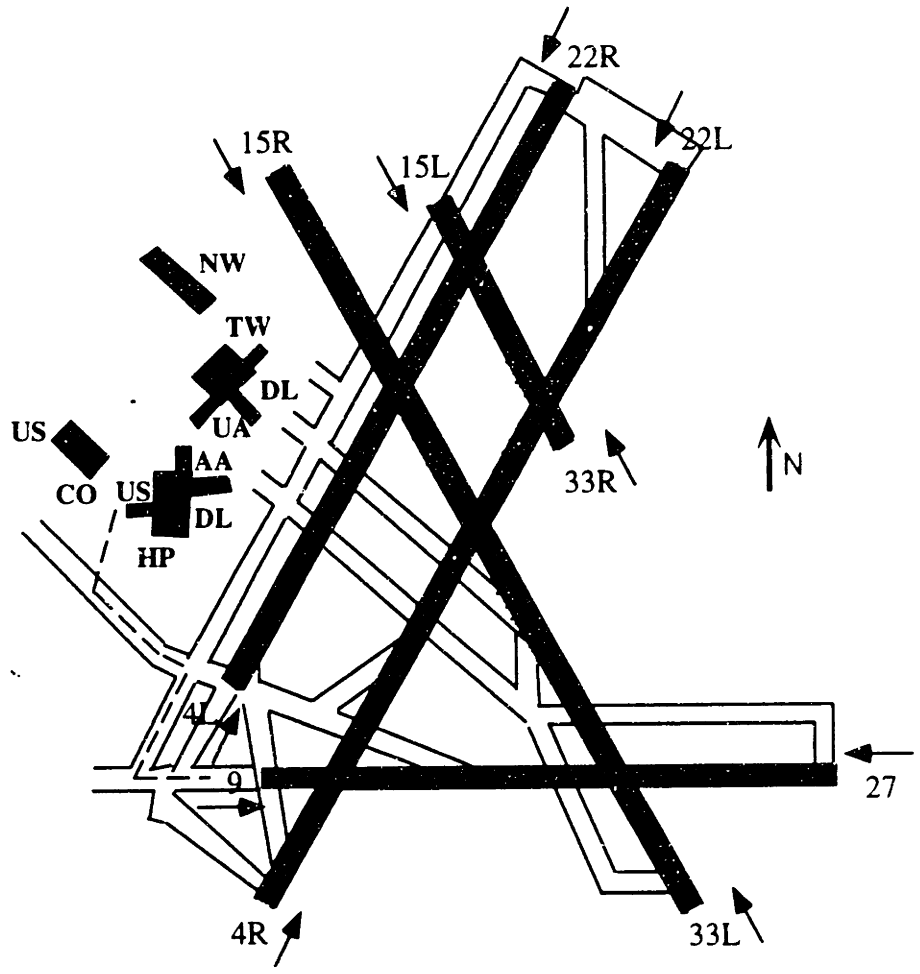


Figure 6-1: Logan Airport runways, taxi-ways, and terminals. USAir departures have a significantly shorter distance from the gate to popular runway 9 than Northwest departures.

model holds constant the presumed impact of each of these conditions.

The differences between static and dynamic linear models extend not only to model form, but also to the approach for finding model parameters. Since the static model parameters Θ are assumed to be constant over time, the value of Θ can be estimated from a large historical database. The next section of this chapter describes how we choose model factors and estimate parameters. Careful attention is paid to outlier detection and residual analysis. If implemented in the ETMS, similar analyses may be performed off-line, when computers and personnel are not devoted to tracking active flights. Subsequent forecasts of roll-out times are made using the coefficient estimate $\hat{\Theta}$ which is obtained from the historical data set. The dynamic models, on the other hand, contain coefficients $\hat{\Theta}_k$ which are fitted recursively by combining information from each newly observed roll-out time with information from

previous observations. New forecasts are produced after each update to $\hat{\boldsymbol{\Theta}}_k$.

For the dynamic model, after k observations, the roll-out time prediction for flight $k+l$ is:

$$\hat{r}_{k+l} = \mathbf{F}'_{k+l}\hat{\boldsymbol{\Theta}}_k \tag{6.2}$$

Forecasts from the static model use $\hat{\boldsymbol{\Theta}}$ in place of $\hat{\boldsymbol{\Theta}}_k$. Most factors in the vector $\mathbf{F}_{k+l}$ will be known in advance, but some of the factors themselves will be forecasts. For example, suppose we wish to forecast the roll-out time of a Continental Airlines flight one-half hour before its scheduled push-back. The model may include indicator variables to adjust the predicted roll-out time so that it corresponds to a Continental flight of the appropriate equipment type and flight duration. Other factors make further adjustments for runway configuration, weather conditions, and the departure queue found by the flight. An important point is that the factors 'carrier' and 'equipment type' can be predicted with great accuracy over any forecast horizon, while runway configuration, weather, and departure queue length are factors which themselves must be modeled and predicted. The values of these factors are uncertain at the time the forecast is generated and may not even be known with certainty when the actual push-back occurs. Forecasts of these factors may be generated by other models, such as the aggregate flow model of the previous chapter or a weather model maintained by the National Weather Service. We would expect that if the forecasts of the factors themselves become more inaccurate as forecast horizons increase, then the roll-out time predictions also suffer. This effect will be explored in the empirical results at the end of the chapter.

The remainder of this chapter is divided into three sections. Section 6.1 describes the development of two static linear models, one containing departure queue lengths and the other using an estimate of departure demand based on the carriers' schedules. Section 6.2 specifies a variety of dynamic models, including exponential smoothing, lookup tables (the procedure currently used by the ETMS), and the dynamic linear model. The potential benefits of combining forecasts from these models is also discussed. Section 6.3 describes empirical tests of these forecasting models. Despite the wide variety of models employed, we achieve only partial success in predicting roll-out times.

Factor	initial coefficient	after outlier removal	after correlation correction	(p-stat)
intercept (min.)	12.6	13.4	13.4	(<.001)
DEMAND (min./aircraft)	0.3	0.3	0.3	(<.001)
US to NE runway	-5.0	-4.5	-4.5	(<.001)
US to SW runway	-2.9	-2.1	-1.9	(<.001)
America West (HP)	-7.0	-4.6	-4.8	(<.001)
Delta (DL)	-2.4	-2.2	-2.2	(<.001)
Northwest (NW)	0.6	1.0	1.0	(.001)
precipitation	2.9	1.8	1.8	(.002)
departure runway 22	3.4	2.5	2.6	(<.001)
departure runway 33 or 27	2.6	2.0	2.1	(<.001)
flight over 3 hours	5.0	2.8	2.2	(<.001)

Table 6.1: Coefficient estimates and p-values for a static linear model which includes 'DEMAND,' an estimate of departure demand derived from the carriers' schedules. All coefficients are in minutes except for the coefficient for DEMAND, which is in min./aircraft. The factor 'US to NE' indicates USAir flights departing on Northeasterly runways 9 or 4. The coefficient 'US to SW' indicates USAir flights departing on Southwesterly runways 22 or 27.

6.1 Static Linear Models

This section describes procedures for model selection and parameter estimation for static linear models of roll-out time. Initially, over 100 factors were considered for the linear models, including weather state, runway configuration, carrier, aircraft weight, time of day, precipitation, length of flight, wind speed and direction, and interactions between many of these factors.

Two measures of airfield congestion were also considered for the models. The first, N_k, was derived from the aggregate flow models of the previous chapter. The aggregate flow models generated estimates of $N(t)$, the number of aircraft rolling out at the beginning of time period t. The quantity N_k is simply an estimate of the number of aircraft rolling when flight k pushes back. That is, if the push-back time of flight k is during time period t, then $N_k = N(t)$. An additional variable, 'DEMAND', was created by counting the number of push-backs *scheduled* during a fifteen-minute window around each flight's actual push-back time. The DEMAND count should approximate the size of the departure queue seen by each push-back without relying on departure queue estimates produced by a forecasting model.

Factor	initial coefficient	after outlier removal	after correlation correction	(p-stat)
intercept (min.)	13.0	13.4	13.4	(<.001)
N_k (min./aircraft)	0.6	0.5	0.5	(<.001)
US to NE runway	-5.8	-5.4	-5.4	(<.001)
US to SW runway	-2.3	-1.8	-1.9	(<.001)
CO to NE runway	-3.5	-2.2	-2.2	(<.001)
CO to SW runway	1.4	2.1	2.1	(<.001)
America West (HP)	-7.4	-4.9	-4.9	(<.001)
Delta (DL)	-1.9	-2.2	-2.3	(<.001)
Northwest (NW)	0.8	0.9	1.0	(.001)
IFR with departure runway 27	1.4	1.3	1.3	(.02)
flight over 3 hours	4.9	2.4	2.2	(<.001)
flight to Newark	-1.3	-1.4	-1.5	(.002)
flight to Laguardia	1.6	0.9	0.8	(.05)
flight to Dallas/Ft. Worth	-0.9	-2.2	-2.2	(<.001)

Table 6.2: Coefficient estimates and p-values for static model which includes N_k, an estimate of airfield congestion. All coefficients are in minutes except for the N_k coefficient, which is in min./aircraft

A model with 100 factors would be cumbersome, and would not necessarily be more effective than a smaller model. This section will describe how we developed two more parsimonious models. The models are shown in the last columns of Tables 6.1 and 6.2, while other colums show intermediate models which will be described below. The primary difference between the linear models shown in the two tables is that the model shown in Table 6.1 contains the variable 'DEMAND,' while the model of Table 6.1 contains the congestion estimate N_k.

Figure 6-2 shows actual roll-out times during two evening hours on March 14, 1991, as well as roll-out time predictions from the static linear model of Table 6.1. The variations in predictions are due to the variations in the factors which apply to each flight. Many of the largest roll-out predictions, such as the third and sixth flights as well as the 'bump' for the eleventh and twelth flights, are flights over three hours long. Table 6.1 indicates that the static linear model adds 2.2 minutes to the predicted roll-out times of each of these flights. Many of the adjacent flights are USAir flights headed to the Northeast runways; these have a roll-out time adjustment of -4.5 minutes. In addition, the plot demonstrates two important shortcomings of the model: the existence of large outliers (the two here are obvious) and autocorrelation in the residuals (note the 'run' of low forecasts between 18:30

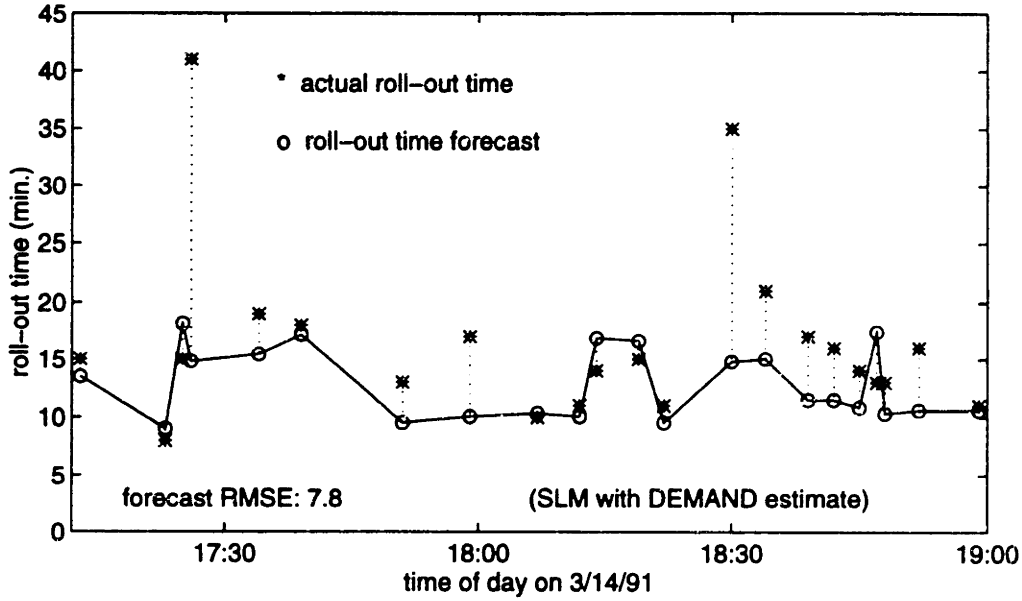


Figure 6-2: Actual roll-out times (*) and roll-out time forecasts (o) from the static linear model with DEMAND estimate. The root mean squared error (RMSE) for these data is 7.8 min. Without the two outliers, the RMSE is 4.8 min.

and 19:00).

Data for model-building were extracted from the matched data set for March and August, 1991. This data set only includes flights by the large carriers and does not contain General Aviation and flights by commuter airlines. The data set includes 44% of all departures from Logan during the sample period. Approximately half of the data are used to build the model and fit its parameters, while the other half are held in reserve for testing. Both of these *training* and *testing* data sets contain representative good-weather and bad-weather days from both March and August. All of the following model-building was performed with the training data.

Model development was an iterative process. In each iteration, the model's parameters were estimated and model residuals were examined. If the residuals indicated any inadequacies in the model, the model was corrected, parameters were re-estimated, and residuals re-examined. This process continued until the model had no apparent shortcomings. The three primary iterations were variable selection, outlier detection, and correction for serial correlation in the residuals. These are discussed in each of the following three sections.

6.1.1 Variable Selection

A stepwise variable selection procedure was adapted to determine the factors to be included in each model. The stepwise procedure relies on simple least-squares regression to select factors and eliminate those that lose significance when others are added [33]. The procedure begins with a single intercept term and no explanatory variables. In its first step the procedure adds the explanatory variable that has the largest correlation with the dependent variable. In the next step it chooses the remaining explanatory variable that has the largest partial correlation with the dependent variable, *given* the variables already included in the model. The partial correlation between roll-out time r_k and factorF_i, given factors $F_1, F_2, \ldots, F_{i-1}$, is the correlation between the residuals when r_k is regressed on $F_1, F_2, \ldots, F_{i-1}$ and the residuals when F_i is regressed on $F_1, F_2, \ldots, F_{i-1}$. The partial correlation is one measure of the benefits of adding F_i to a model which includes $F_1, F_2, \ldots, F_{i-1}$.

Additional independent variables are added in the same manner. At each step, partial F statistics are calculated for all variables in the model, and variables with an F statistic below some threshold are dropped. For this experiment, the F-test used an F statistic of 2, for a p-value of approximately 0.15. If any variable's F statistic has a p-value above 0.15, it is dropped from the model. This procedure continues until any new independent variable added to the model fails to pass over the threshold F statistic.

Stepwise regression has many potential pitfalls but there is one very obvious difficulty with this procedure: the significance calculated for each F-test is unreliable. The F-tests are statistically valid if model errors are independent and normally distributed with a mean of zero and a constant variance. These conditions do not hold for these data. However, we will see that initial model selection will not hinge on the values of the F-tests. A few factors will clearly belong in the model, many will not, and the ones in between will not make a significant difference in terms of model performance.

The stepwise procedure was applied separately to the March and August data. Figure 6-3 show the increasing R^2 values for the August data as factors are added to each model. The bottom line shows the progression of R^2 values when the congestion estimate N_k is not included as a factor. In this case, no one factor is the primary contributor to the model's explanatory power. The first factor added, DEMAND, leads to an increase in R^2 of only 0.04. Subsequent additions to the model lead to gains that slowly taper off as the model reaches a limiting R^2 of about 0.27. The top line shows the sequence of R^2 values when N_k

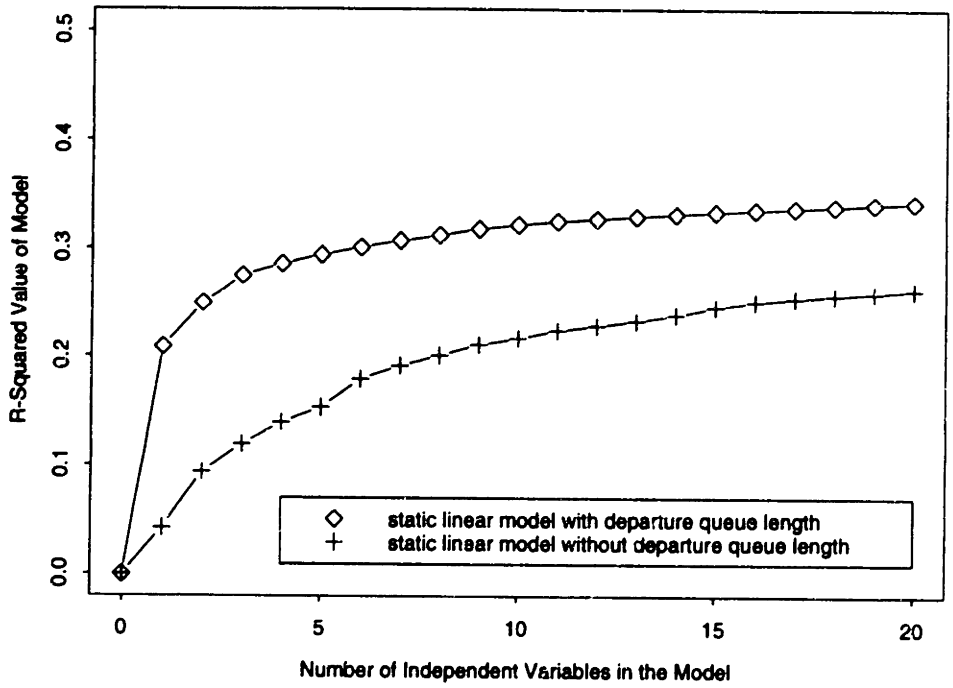


Figure 6-3: Value of R^2 with August data as factors are added to the model. In the top line the first factor added is N_k, an estimate of the length of the departure queue.

is added to the model. The first factor added is N_k itself, and this immediately increases the R^2 value to 0.2. The second factor, an indicator variable that the flight is operated by USAir and is headed for a nearby departure runway, has a marginal benefit of about 0.04. Additional factors have an even smaller effect. Long departure queues are not only highly correlated with long roll-out times, but the departure queue length is also correlated with other likely explanatory variables, such as weather, low-capacity runway configurations, and heavy demand. Alone, it is a good substitute for many other factors.

Despite the preeminence of N_k, the large number of observed flights enabled the procedure to identify over 30 significant independent variables in each month for both models. Although all of these variables pass the F-test, most variables added late in the stepwise procedure had little impact on the model's explanatory power. Therefore, in order to find more parsimonious models, the following ad-hoc rules for choosing factors were applied to each model:

1. Any variable among the first five selected by the stepwise regression procedure in *either* March or August is included in the model;

2. Any variable among the first 30 in *both* March and August is included in the model.

These rules ensure that the most significant effects in both months are represented in the models. Table 6.1 lists the factors and coefficient values for the static model without departure queue estimates. Table 6.2 lists the factors and coefficient values for the static model with N_k, the departure queue estimates. The coefficient values listed in the tables were obtained by fitting the data to the model by ordinary least-squares (OLS), using the entire training data set of March and August data. The R^2 values of the models with, and without, queue length estimates are 0.24 and 0.16, respectively (the values differ from those shown in Figure 6-3 since that figure was generated from August data only, while these R^2 values were calculated from the entire training data set).

It is a valuable exercise to examine the factors selected for the models and their coefficient values. Terms for departure demand and departure queues are not surprising, nor are terms which reflect airport geography. The locations of USAir (US), Continental (CO) and America West (HP) gates in relation to the most common departure runways, 4 and 9, explain the negative coefficient values for these carriers (see Figure 6-1). Geography also explains Northwest's comparative disadvantage. The additional roll-out time due to precipitation is not unexpected. While taxiing, ground controllers use surface radar to aid aircraft, but visual navigation is the pilot's responsibility. Heavy rain reduces visibility and can slow surface throughput. The significance of flights to Newark, LaGuardia, and Dallas/Ft. Worth is a bit surprising. It is possible that flights to these destinations have gate locations close to the taxi-ways, or uncongested apron areas surrounding the gates.

However valid the models seem, the explanatory power of the models, as measured by the R^2 statistic, is quite low when one considers the number of factors included in each model. A look at the model residuals may give us some idea about what we may be missing.

6.1.2 Outlier Detection and Removal

During the two-hour period shown in Figure 6-2 there are two roll-out times with significantly larger residuals than the others. These roll-outs lasted 41 and 35 minutes, and were both generated by Pan Am flights. The first was destined for JFK and the other for Miami. Over all roll-outs shown in Figure 6-2, the root-mean squared error (RMSE) of the model is 7.8 min. Without these two longest roll-outs, the RMSE drops to 4.8 min. Successful prediction of these two would have greatly improved the model's performance, but it seems

unlikely that these long delays could have been anticipated given the information in the data set. Pan Am flights before and after these outliers did not suffer significant delays, nor did flights to the same destinations. In fact, no pattern in the existing data adaquately explains these long delays.

The residuals produced by these Pan Am flights were not unusual. Figure 6-4 displays the largest 200 residuals from the static linear model with departure queue estimates, ordered from largest to smallest in absolute value. The largest residual is almost 90 min. If the true model errors ϵ_k were distributed as a normal distribution, then the *standardized residuals* e_k^s from our OLS model should be distributed as standard normals. The standardized residuals are the residuals, e_k, divided by their standard errors:

$$e_k^s = \frac{e_k}{s\sqrt{1-h_{kk}}} \tag{6.3}$$

where

$$e_k = \left(r_k - \mathbf{F}_k'\hat{\mathbf{\Theta}}\right) \tag{6.4}$$

$$s^2 = \frac{\sum_{k=1}^{N} e_k^2}{N-p-1} \tag{6.5}$$

$$h_{kk} = (\mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}\mathbf{F})_{kk} \tag{6.6}$$

Here, p is the number of explanatory variables, and the quantity h_{kk} is a measure of the distance between the explanatory vector F_k and the average explanatory vector in the entire data set. If the 'distance' h_{kk} is large, then errors in $\hat{\mathbf{\Theta}}$ may be exagerrated and we would expect the residual of flight k to have a comparatively large variance.

The Quantile-Quantile (Q-Q) plots in figure 6-5 compare the distribution of the standardized residuals plotted against a sample from a standard normal distribution. Each point represents a residual, with the value of the standardized residual as the abscissa. The ordinate is the value from the standard normal sample which has the same cumulative percentile in the standard normal sample distribution. If (e_k^s, n_k) are the coordinates of point k, then prob$(e \leq e_k^s)$ = prob$(n \leq n_k)$, where e and n are distributed as the standardized residuals and the standard normal, respectively. Samples with similar distributions approximate a straight line; the residuals which trail off to the right in the top graph demonstrate that the distribution of residuals has a much longer right tail than a normal distribution. Besides invalidating the standard statistical tests (such as the F-tests for the significance

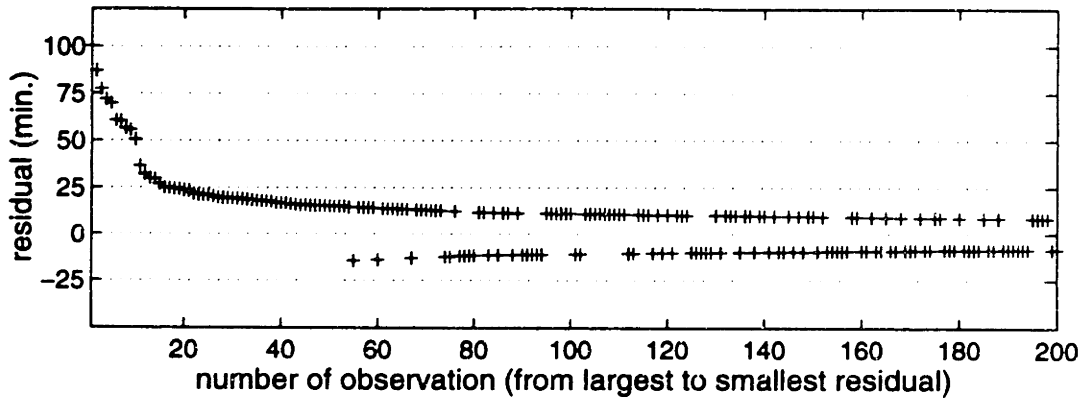


Figure 6-4: Residuals from the original static model.

of coefficients), these outliers have a significant impact on both model form and fit. Large outliers have an unreasonable influence on the values of model parameters while diminishing the R^2 value.

If the outliers suggest shortcomings in the model, then the model should be corrected and re-fitted. Graphical and analytical analysis of the residuals found one pattern, serial correlation, which will be corrected in the next section. However, no general explanation for the largest residuals was found. The largest residuals represent unique or unusual conditions that are not repeated in a predictable way. The outliers may be errors in data transcription or they may have been caused by isolated delays due to passengers, crew, or the aircraft. Whatever their cause, true outliers should be removed before the model is fitted again. But how many residuals should be considered 'outliers' and removed? In this large data set, the removal of any observation with a large residual will improve the R^2 value. However, removing true outliers should produce another benefit. Once the outlier and its insidious effect on the model coefficients is removed, the R^2 value of a model fitted to the remaining points should be superior to the R^2 value of the model fitted on the original data set.

With this observation in mind, we ranked observations in the order of the absolute values of their standardized residuals, largest to smallest. We then began removing observations from the data set in that order. After each observation was removed, a new R^2 value was calculated using the original parameters of the static linear model. In addition, a new 're-fitted' model was found by ordinary least squares applied to the remaining observations, and a re-fitted R^2 was calculated. As the most unrepresentative outliers are removed, the marginal gain in R^2 should be greater for the re-fitted model than for the original model. However, once these observations are gone and more representative data begin to

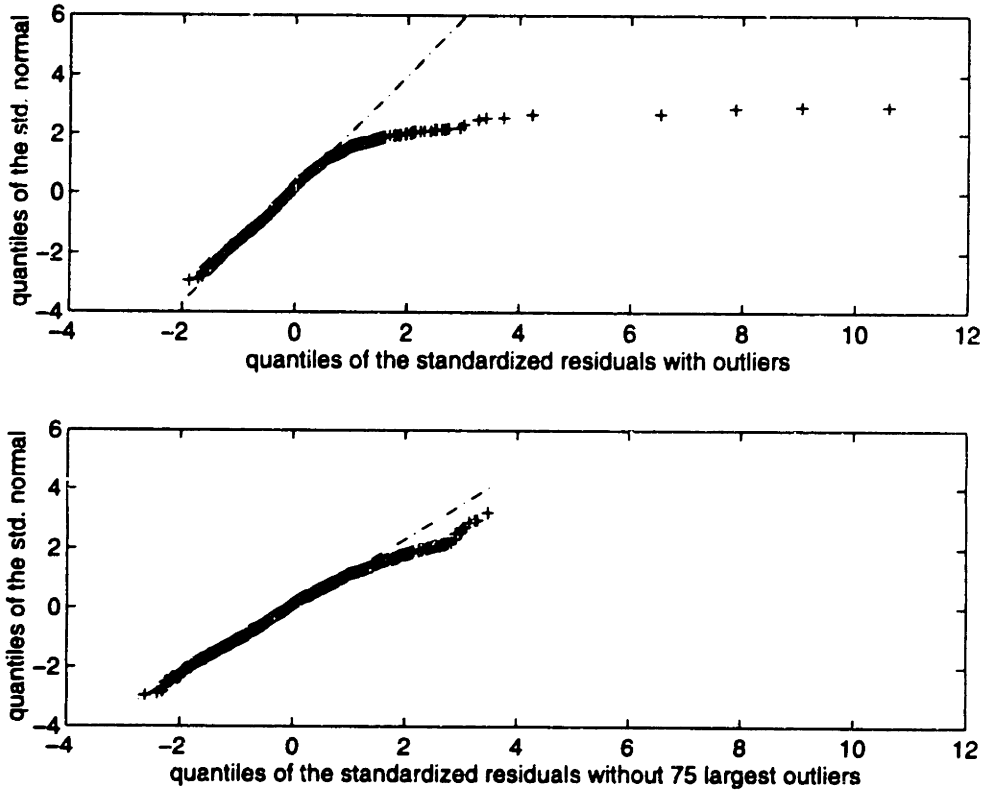


Figure 6-5: Q-Q plots of standardized residuals against samples from a unit normal for models with and without outliers. The x-axis represents values of the standardized residuals, the y-axis values from the standard normal distribution. Each point has x and y-values with the same percentiles in their respective distributions.

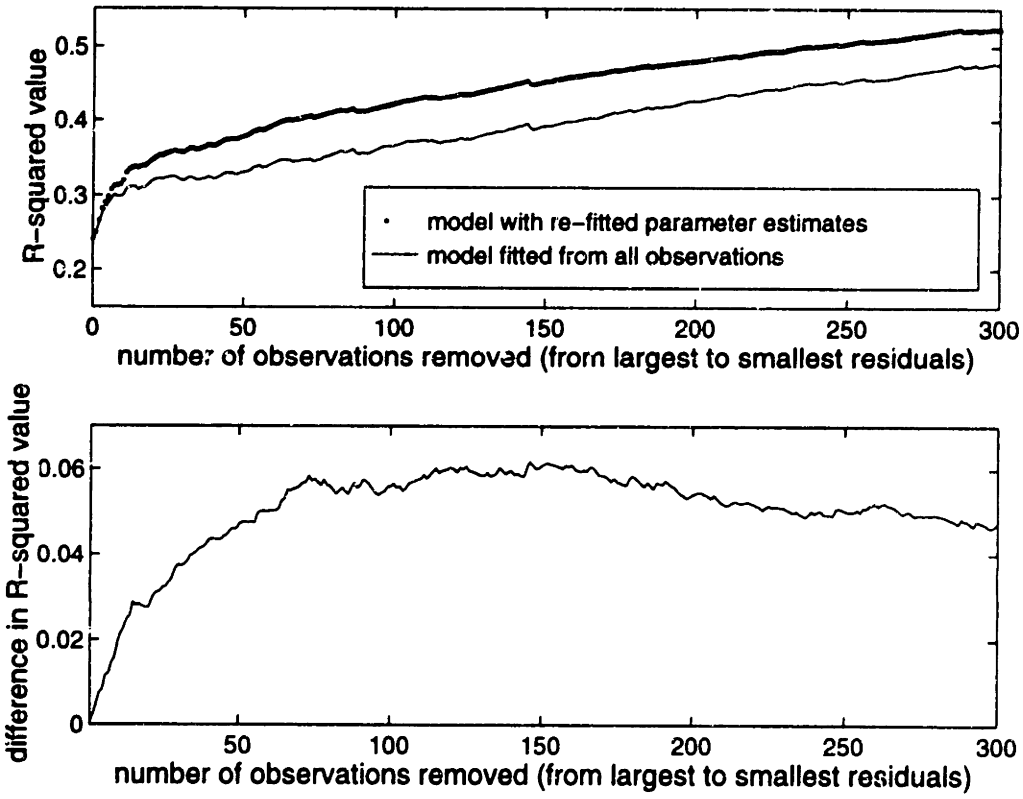


Figure 6-6: Increasing value of R^2 as outliers are removed from the data set (top). At first, the R^2 values increase more quickly if the model is re-fitted; differences in R^2 between original and re-fitted models are shown in the bottom plot.

be deleted, the re-fitted models should no longer be superior to the original model on the margin. Figure 6-6 shows the results of this procedure. The top plot shows that the R^2 values of both the original and re-fitted models improve as outliers are removed. However, the gains of the re-fitted model are at first greater than the gains of the original model. The bottom plot shows the difference between the R^2 values for these two models. The marginal benefits of outlier removal are larger for the re-fitted model as the first $75 - 100$ outliers are removed. These results indicated that we should consider as outliers the observations which produced the 75 largest standardized residuals. After removing these observations, we calculated a new static linear model using ordinary least squares. The results are seen in the second columns of Tables 6.1 and 6.2, and the residuals are shown in the bottom Q-Q plot of Figure 6-5. The R^2 values of the re-fitted models (without residuals) are 0.41 and 0.29 for the models with and without departure queue length estimates, respectively. However, even after these outliers have been removed, the residuals show one more pathology: serial correlation.

6.1.3 Correcting for Serial Correlation in the Residuals

During the two-hour period in Figure 6-2, there appears to be a positive correlation in the residuals. During this sample period, almost all forecasts are too low, and this trend is especially clear between 18:30 and19:00 (for the sample after outlier removal, the mean residual is zero). Some of these variations in roll-out times may be caused by temporary conditions which apply to flights which are close together in time. Since the roll-out time observations r_k are ordered by take-off time, model errors ϵ_k may be serially correlated. The observed model residuals e_k do display significant autocorrelation. An estimate of the first-order autocorrelation is:

$$\hat{\rho} = \frac{\sum_{k=2}^{N} e_k e_{k-1}}{\sum_{k=1}^{N} e_k^2} \tag{6.7}$$

For both static linear models, $\hat{\rho} = 0.18$. A Durbin-Watson test found a significant autocorrelation (p-value $< .0001$). A simple model for the residuals is the first order autoregressive (AR(1)) process :

$$\epsilon_k = \rho \epsilon_{k-1} + u_k \quad \text{for } k = 2 \ldots N \tag{6.8}$$

$$\epsilon_1 = u_1 \tag{6.9}$$

where $|\rho| < 1$ and the u_k are independent and identically distributed with zero mean and a constant variance. A consistent estimate of ρ is $\hat{\rho} = 0.18$. In fact, residuals from *this* model show little significant correlation, indicating that the AR(1) process is a reasonable model for the residuals.

A simple transformation of the original data allows us to use ordinary least squares (OLS) to find new estimates for $\boldsymbol{\Theta}$, given the first-order correlation in the residuals:

$$r_k - \rho r_{k-1} = (F_k - \rho F_{k-1})\boldsymbol{\Theta} + \upsilon_{\mathbf{k}} \tag{6.10}$$

or, after redefining the left and right-hand sides:

$$r_k^* = F_k^{*'}\boldsymbol{\Theta} + \upsilon_{\mathbf{k}} \tag{6.11}$$

The transformed observations and factors r_k^* and F_k^* were estimated using $\hat{\rho}$. Ordinary least squares was applied to these data to find another parameter estimate for $\boldsymbol{\Theta}$. Results and p-values are shown in the third and fourth columns of Tables 6.1 and 6.2. There was little change in the coefficient values with the elimination of autocorrelation. The p-values show that all selected explanatory variables are highly significant.

The existence of autocorrelation in the residuals suggests that models of roll-out times which contain an explicit temporal component may lead to improvements in model accuracy. Such models, which attempt to capture local changes in roll-out times, will be described in the next section.

6.2 Dynamic Models

The linear model developed in the previous section is a static model, with coefficients $\boldsymbol{\Theta}$ obtained via ordinary least squares (OLS) regression applied *en masse* to historical data. The OLS procedure gives each observation equal weight, but we would expect recent observations to be more informative than roll-out times in the distant past. This suspicion is confirmed by the serial correlation in the residuals of the static linear model. The roll-out time forecasts in Figure 6-2 provide an additional example: the forecasts are consistently too

low, but the model does not adapt. The model may minimize the sum of squared residuals over the entire data set, but it cannot adapt to what may be temporary conditions on the airfield.

Improvements in forecasting accuracy may be achieved by more responsive forecasting procedures. This section presents four such procedures: exponential smoothing, a lookup table, exponentially weighted linear regression, and dynamic linear models. Throughout the section we will see how the procedures perform with the two-hour period shown in Figure 6-2. We will also use the following simple example to demonstrate the mechanics of each technique.

EXAMPLE 6.1 Suppose that the forecaster observes a sequence of roll-out times, but that the only significant factor that is available to the forecaster is the weather condition at the departure airport. In addition, suppose that weather is classified into one of two states: good and bad. During the period of observation there is a stretch of bad weather, a long period of good weather, and then a return to bad weather. Our forecaster observes a sequence of m roll-out times under bad weather, a sequence of n roll-out times under good weather, and then a single roll-out under bad weather. In a linear model, a bad weather state is represented by an indicator variable, with each good weather roll-out assigned a 0 and each bad weather roll-out represented by a 1. The sequence of factors is:

$$\left[\underbrace{11\ldots1}_{m} \quad \underbrace{00\ldots0}_{n} \quad 1 \right]$$

and there are $m + n + k$ roll-out time observations, $r_1, r_2, \ldots, r_{m+n+k}$.

6.2.1 Exponential Smoothing and the Lookup Table

At any time, a simple forecast for the next roll-out might be the average of all past roll-out times. However, we would expect recent roll-out times to be more relevant than observations taken further in the past, for it is likely that the next departure will experience conditions similar to those experienced by recent flights. Therefore, $\hat{r}_{k+1}$, the predicted roll-out time of flight $k + 1$ given information up to the kth roll-out, will be an exponentially weighted (or 'smoothed') average of previous roll-out times:

$$\hat{r}_{k+1} = C_k(1-f)\sum_{i=1}^{k} f^{k-i} r_i \tag{6.12}$$

$$C_k = (1 - f^{k+1})^{-1} \tag{6.13}$$
$$0 < f < 1$$

The $(k-l)$th observation is weighted by f^l. Since the normalizing constant C_k approaches 1 as k becomes large, for large k we have the recursive expression:

$$\hat{r}_{k+1} \approx (1 - f)r_k + f\hat{r}_k \tag{6.14}$$

Our choice of the constant 'fade factor' f determines how sensitive the average is to changes in observed roll-out times. For example, experiments with the testing data set from Logan Airport indicate that setting f equal to 0.9 produces the greatest forecast accuracy. With this choice of f, 10% of each forecast is contributed by the most recent observation and 90% from the weighted average of all other observations. The top plot in Figure 6-7 shows the forecasts made by the exponential smoother during the two-hour period on March 14. The forecasts change very slowly over time, with relatively large 'jumps' only immediately after the two outliers. It is surprising that this simple procedure has a RMSE close to that of the static linear model shown in Figure 6-2 (7.8 min. for the linear model, 7.9 min. for the exponential smoother). Of course, this is a small sample, and we will be comparing the procedures over a much larger set of data later in this chapter. However, it does seem as if the adaptability of the smoother gives it one advantage over the static procedure.

However, the static linear model does take advantage of the many factors that are known to affect roll-out times. For example, the model generates different forecasts for flights departing in good and bad weather, since historical roll-out times are significantly different under these conditions. One extension of the exponential smoother is to compile a different average from each set of flights that experienced each set of operating conditions. For example, roll-out times of heavy aircraft in bad weather may be averaged, and this average may be used to forecast roll-out times of heavy aircraft in bad weather. These running averages are stored in a *lookup table*. Such a system has been developed at the Volpe center and is now being implemented for the ETMS [37].

EXAMPLE 6.1 (continued). Our lookup table has two averages, or cells: one for bad weather ($\hat{r}_k^b$) and one for good weather ($\hat{r}_k^g$). After observation $m + n$, the bad weather average has been

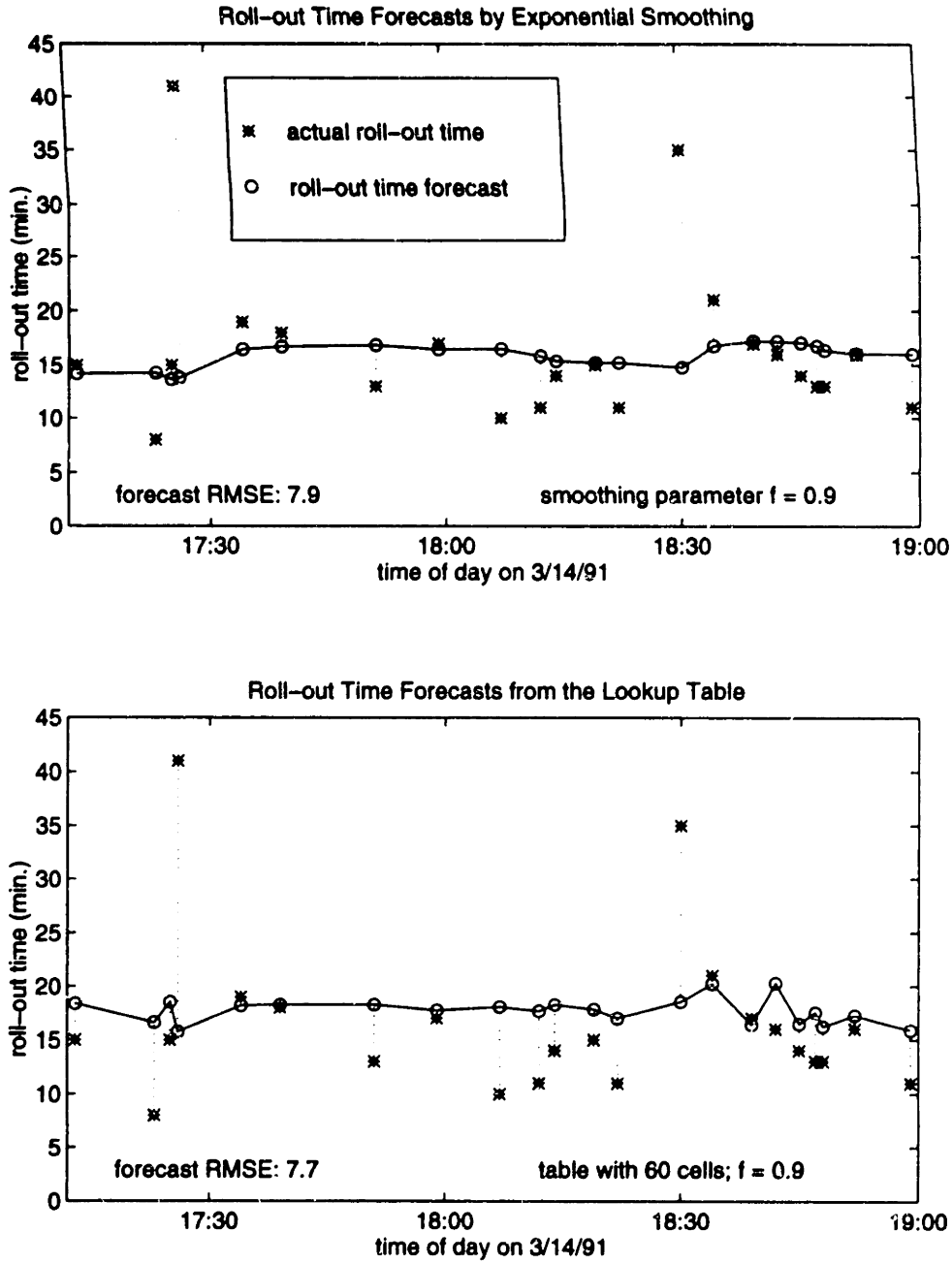


Figure 6-7: Actual roll-out times (*) and roll-out time forecasts (o) from exponential smoothing and the lookup table. The exponential smoother gives 10% of its weight to the latest observation: $\hat{r}_{k+1} = (.1)r_k + (.9)\hat{r}_k$. The lookup table uses a similar smoother, but divides the observations among 60 cells categorized by weather, flight duration, and time of day.

compiled from the first m flights and the good weather average from the next n flights:

$$\hat{r}^b_{m+n+1} \approx \sum_{k=1}^{m}(1-f)f^{m-k}r_k \tag{6.15}$$

$$\hat{r}^g_{m+n+1} \approx \sum_{k=m+1}^{m+n}(1-f)f^{m+n-k}r_k \tag{6.16}$$

Since flight $m+n+1$ will operate under bad weather, we will use $\hat{r}^b_{m+n+1}$ as our forecast. Note that this forecast is calculated from observations taken over n flights ago.

The bottom plot of Figure 6-7 provides another example. The forecasts in this plot are generated from a lookup table indexed on three categories: weather state, flight duration and time of day. There are four weather states (VFR, MVFR, IFR, LIFR), five categories of flight duration ($0-1, 1-2, 2-3, 3-4$, and >4 hours), and three time-of-day categories (morning, afternoon, and evening) creating a lookup table with sixty cells. This is similar to the table implemented by the ETMS, although the ETMS has a day-of-week category as well. During the sample period shown in Figure 6-7, all flights are evening flights under IFR, so flights are only distinguished in this lookup table by their length. This lookup table does differentiate between flights to a greater extent than the exponential smoother, but forecasts do not vary as widely as those from the the static linear model. However, the performance of the lookup table for this sample is comparatively good: a RMSE of 7.7 min., slightly lower than that of the SLM. Finally, none of the models presented so far are successful in predicting the two outliers shown in the figure.

A closer look at these examples demonstrates a potential pitfall of the lookup table: there may be large gaps between predictions from the same cell. For example, the first flight in Figure 6-7 is a two-hour flight, and the next flight from the same cell departs at 18:07, almost one hour later. This effect is also seen in example 6.1, above, when the forecast for a bad-weather flight did not include information from the most recent flights. These gaps will not affect model performance if conditions change slowly and the data are not distributed among many cells. However, the problem may be exacerbated when more factors are used to forecast roll-out times. For example, if we were to compile averages under 4 weather states, 7 runway configurations, 10 carriers, and 6 levels of departure demand, we would need 1680 separate averages, or cells, in the lookup table. As we add factors we break the data into more cells, and any one forecast is based on fewer recent observations.

One alternative is to combine the effects of all factors, so that all observations can be used to inform the same model.

6.2.2 Recursive Regression and Exponentially Weighted Regression

The static linear model describes roll-out time as the sum of the relevant factors, each multiplied by a coefficient. The coefficients represent the increase in roll-out time when there is a unit increase in the associated factor. As we saw in the previous section, the linear model for a single observed roll-out r_i is:

$$r_i = \mathbf{F}_i'\boldsymbol{\Theta} + \epsilon_i \tag{6.17}$$

$$\tag{6.18}$$

A series of k observations can be expressed in matrix-vector form:

$$\mathbf{r}_k = \underline{\mathbf{F}}_k'\boldsymbol{\Theta} + \boldsymbol{\epsilon}_k \tag{6.19}$$

where

$$\underline{\mathbf{F}}_k = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k] \qquad \mathbf{r}_k = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} \qquad \boldsymbol{\epsilon}_k = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}$$

Note that $\underline{\mathbf{F}}_k$ is a matrix built from the vectors $\mathbf{F}_i$, $i = 1 \ldots k$.

EXAMPLE 6.1 (continued). Our model consists of the intercept, θ_0 and a single factor, the weather. The influence of the weather is represented by θ_1, while the presence of this factor at observation k is indicated by F_{k1}. F_{k1} is 0 during good weather and 1 during bad weather. The indicator for the intercept, F_{k0}, is always 1. After observing $m + n + 1$ departures, our factor and coefficient matrices are:

$$\underline{\mathbf{F}}_{m+n+1} = \begin{bmatrix} 111\ldots1 & 111\ldots1 & 1 \\ \underbrace{111\ldots1}_{m} & \underbrace{000\ldots0}_{n} & 1 \end{bmatrix} \qquad \boldsymbol{\Theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \tag{6.20}$$

After observing some number k of roll-out times, we should be able to estimate the

unknown parameters in $\boldsymbol{\Theta}$. In the previous section on static linear regression, we used ordinary least-squares estimates, which minimize the sum of squared residuals:

$$\begin{aligned} \hat{\boldsymbol{\Theta}}_k &= \min_{\boldsymbol{\Theta}} \left[\sum_{i=1}^{k} (r_k - \mathbf{F}_k' \boldsymbol{\Theta})^2 \right] \\ &= \min_{\boldsymbol{\Theta}} \left[\sum_{i=1}^{k} e_i^2 \right] \end{aligned} \tag{6.21}$$

The well-known formula for the solution to the minimization problem is

$$\hat{\boldsymbol{\Theta}}_k = \left(\underline{\mathbf{F}}_k \underline{\mathbf{F}}_k'\right)^{-1} \underline{\mathbf{F}}_k \mathbf{r}_k \tag{6.22}$$

A necessary and sufficient condition for the invertibility of $\underline{\mathbf{F}}_k \underline{\mathbf{F}}_k'$ and the uniqueness of $\hat{\boldsymbol{\Theta}}_k$ is that $\underline{\mathbf{F}}_k'$ have full column rank.

In this section we are most interested in efficiently incorporating new information as it arrives. After we have observed k roll-outs, found $\hat{\boldsymbol{\Theta}}_k$, and then observed an additional departure, we might be tempted to append the additional observation to our matrices and resolve equation 6.22 for $\hat{\boldsymbol{\Theta}}_{k+1}$. However, the matrix operations for this calculation may become prohibitively expensive as k becomes large. Since we have already found $\hat{\boldsymbol{\Theta}}_k$, the following recursive equation is much more efficient:

$$\hat{\boldsymbol{\Theta}}_{k+1} = \hat{\boldsymbol{\Theta}}_k + \mathbf{P}_{k+1}^{-1} \mathbf{F}_{k+1} \left(r_{k+1} - \mathbf{F}_{k+1}' \hat{\boldsymbol{\Theta}}_k \right) \tag{6.23}$$

$$\mathbf{P}_{k+1} = \mathbf{P}_k + \mathbf{F}_{k+1} \mathbf{F}_{k+1}' \tag{6.24}$$

$$\mathbf{P}_0 = \mathbf{0}$$

The difference $\left(r_{k+1} - \mathbf{F}_{k+1}' \hat{\boldsymbol{\Theta}}_k \right)$ is the prediction error, while $\mathbf{P}_{k+1}^{-1} \mathbf{F}_{k+1}$ is a vector which weights this error before it is added to the previous estimate of $\boldsymbol{\Theta}$.

The least-squares estimation procedure is one of the most popular forecasting tools in practice, but the results of the previous sections have demonstrated one potential drawback: the OLS criterium (6.21) gives as much weight to recent observations as to those in the distant past. In practice, the ETMS forecasting system may run for years and will observe hundreds of thousands of roll-out times at each airport. Temporary shifts in the level of the roll-out times, like those shown in Figure 6-2, will not be reflected in the model.

One method for reducing the influence of flights that occurred in the distant past is to discount their residuals in equation (6.21). We redefine the optimal estimate for Θ after k observations as:

$$\begin{aligned} \hat{\Theta}_k &= \min_{\Theta} \left(\sum_{i=1}^{k} f^{k-i} e_i \right) \\ \text{where} \quad & 0 < f \leq 1 \end{aligned} \tag{6.25}$$

The residual l observations before k receive a weight of f^l, so that the influence of an observation decreases exponentially as it recedes into the past. Note that if $f = 1$, this is equivalent to the standard least-squares criterion.

The solution to equation 6.25 is also similar to the least-squares solution in the last section. Using our definitions of $\underline{\mathbf{F}}_k$ and $\mathbf{r}_k$ above,

$$\begin{aligned} \hat{\Theta}_k &= \left(\underline{\mathbf{F}}_k \mathbf{S}_k \underline{\mathbf{F}}_k'\right)^{-1} \underline{\mathbf{F}}_k \mathbf{S}_k \mathbf{r}_k \\ \mathbf{S}_k &= \operatorname{diag}\left(f^{k-1}, f^{k-2}, \ldots, f, 1\right) \end{aligned} \tag{6.26}$$

As in the last section, once $\underline{\mathbf{F}}_k \mathbf{S}_k \underline{\mathbf{F}}_k'$ is nonsingular, $\hat{\Theta}_k$ can be uniquely determined. Given this estimate we may continue with the recursion:

$$\hat{\Theta}_{k+1} = \hat{\Theta}_k + \mathbf{P}_{k+1}^{-1} \mathbf{F}_{k+1} \left(r_{k+1} - \mathbf{F}_{k+1}' \hat{\Theta}_k \right) \tag{6.27}$$

$$\mathbf{P}_{k+1} = f\mathbf{P}_k + \mathbf{F}_{k+1} \mathbf{F}_{k+1}' \tag{6.28}$$

$$\mathbf{P}_0 = \mathbf{0}$$

With a fade factor $f < 1$, this estimate is more sensitive to temporary changes in roll-out times than standard linear regression, especially when k grows large. However, this sensitivity comes at a price, for the procedure is unstable if there is a long sequence of observations which do not supply any information about a particular parameter. During such a period, the current estimate of that parameter is quickly discounted, or forgotten, and later estimates do not use old information that may still be valuable. The example illustrates this phenomenon:

EXAMPLE 6.1 (continued). Suppose we wish to estimate $\Theta = (\theta_1, \theta_2)'$ after $N = m + n + 1$ observations. There were m observations under bad weather, n observations under good weather,

and the most recent observation is taken after a return to bad weather. We are interested in what happens when there is a long sequence of observations which do not supply information about the bad weather parameter, θ_2.

It is not hard to show that $\hat{\theta}_{N,2} \approx r_N - \hat{\theta}_{N,1}$ as n becomes large (explicit calculation of $\left(\hat{\theta}_{N,1}, \hat{\theta}_{N,2}\right)$ is performed in Appendix 6A). As the time since the last bad-weather observation increases, previous estimates of the bad-weather coefficient receive less and less weight. If the stretch of good weather is long enough, the estimate $\hat{\theta}_{N,2}$ is based solely on the single bad-weather observation r_N. All previous information about bad weather is ignored.

The source of the problem illustrated in the example is the undiscriminating effect of the fade factor f. Equation 6.25 tells us that observations are discounted exponentially, no matter which factors are active for each observation. Some coefficients, like the intercept θ_1, are updated frequently by new information, and for these coefficients a fade factor less than 1 is appropriate. However, the factors associated with bad weather or rare runway configurations may appear infrequently, and a fade factor near or equal to 1 is needed to maintain the little information that is available. Exponentially weighted regression simply applies the same fade factor to all coefficients.

6.2.3 Dynamic Linear Models

In the terminology of systems theory, the parameter vector $\boldsymbol{\Theta}_k$ in Example 6.1 is *unobservable* during the n good weather observations. If these observations are the only available information, there is no unique best estimate of the bad-weather parameter θ_2 without prior information. The dynamic linear model assimilates this prior information, and the estimation procedures derived from the model can control the degradation of information when the system is unobservable.

The goal of each estimation procedure described above has been to estimate the current true value of the unknown vector of coefficients, $\boldsymbol{\Theta}$. We will call this vector the *state vector.* Its true value may change over time, and in our notation the true value when observation k is taken is $\boldsymbol{\Theta}_k$. The estimation equations will determine a best estimate for $\boldsymbol{\Theta}_k$ given all information up to and including the kth observed roll-out. The following system and update equations for the DLM are the matrix forms of the Kalman filter, which was introduced in Chapter 5. For the aggregate flow models we were estimating the current level of the

take-off capacity $\mu(t)$, while here we are estimating the value of the coefficient vector $\boldsymbol{\Theta}_k$.

First we specify a probability model for the evolution of $\boldsymbol{\Theta}_k$ over time. Assume that $\boldsymbol{\Theta}_k$ and roll-out times r_k are governed by the following equations:

$$r_k = \mathbf{F}_k'\boldsymbol{\Theta}_k + v_k \tag{6.29}$$

$$\boldsymbol{\Theta}_k = \boldsymbol{\Theta}_{k-1} + \mathbf{w}_k \tag{6.30}$$

$$\text{where } v_k \sim [0, V] \text{ and } \mathbf{w}_k \sim [\mathbf{0}, \mathbf{W}]$$

Also assume that both v_k and $\mathbf{w}_k$ are uncorrelated white noise processes with mean zero and constant variance. The model implies that each observation r_k is a linear combination of the components of $\boldsymbol{\Theta}_k$ perturbed by a random variable v_k. Meanwhile, the true value of the state vector $\boldsymbol{\Theta}_k$ evolves according to a random walk with variance $\mathbf{W}$.

The DLM is not so much an exact model of the roll-out process, but a method for detecting and estimating changes in roll-out conditions. The assumption that the effects of weather or other factors follow a random walk is implausible, but the advantages of the model lie in its implementation. The model leads naturally to estimation equations that overcome the difficulties of the lookup table and exponentially weighted regression. The optimal estimate adapts to local changes in the state vector, but by selectively adjusting the covariance matrix $\mathbf{W}$, we may control the loss of information that may result from this adaptability.

We now present the update equations for the estimates. Define the optimal estimate $\hat{\boldsymbol{\Theta}}_k$ to be the linear least squares estimate, the estimate over all linear functions of the input r_k that minimizes the expected squared error. If V and $\mathbf{W}$ are normally distributed then the estimate derived here is also the Bayes' Least Squares estimates. This estimate minimizes the expected squared error over all functions of the input. Assume that before any observations are made we construct prior estimate $\hat{\boldsymbol{\Theta}}_0$ and covariance matrix $\mathbf{P}_0$. The following recursive equations produce optimal estimates from each additional observation:

$$\hat{\boldsymbol{\Theta}}_k = \hat{\boldsymbol{\Theta}}_{k-1} + \mathbf{G}_k\left(r_k - \mathbf{F}_k'\hat{\boldsymbol{\Theta}}_{k-1}\right) \tag{6.31}$$

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{G}_k\mathbf{G}_k'Q_k \tag{6.32}$$

where

$$Q_k = \mathbf{F}_k'(\mathbf{P}_{k-1} + \mathbf{W})\mathbf{F}_k + V \tag{6.33}$$

$$\mathbf{G}_k = (\mathbf{P}_{k-1} + \mathbf{W})\mathbf{F}_k/Q_k \tag{6.34}$$

The update equations have an intuitive structure. The revised estimate of $\boldsymbol{\Theta}_k$ is equal to the previous estimate, adjusted by an amount which is proportional to $\left(r_k - \mathbf{F}_k'\hat{\boldsymbol{\Theta}}_{k-1}\right)$, the one-step forecast error for the old estimate. The vector $\mathbf{G}_k$ is called the *gain* of the filter, and governs the weight given to new observations. With each observation, the covariance matrix of the estimate, $\mathbf{P}_k$, is reduced by an amount calculated from the gain.

For a given set of factors $\mathbf{F}_k$, the size of the gain is determined by both the observation variance V and the system covariance matrix $\mathbf{W}$. The gain is analogous to the fade factor f in exponential smoothing and exponentially weighted regression. If $\mathbf{W}$ is large and V is small, $\boldsymbol{\Theta}_k$ would be expected to change relatively quickly over time and the observation r_k should be given much weight. On the other hand, if $\mathbf{W}$ is small and V large, a long history of observations should be used to estimate $\boldsymbol{\Theta}_k$. In fact, if $\mathbf{W} = 0$ and we have no prior information about $\boldsymbol{\Theta}_k$, then the Kalman Filter is equivalent to the static recursive regression in equations 6.23 and 6.24. The advantage of the DLM over previous algorithms, however, is that a different variance may be specified for each coefficient in $\boldsymbol{\Theta}_k$, while the fade factor f was applied indiscriminately to all coefficients.

The estimation equations for the DLM were first proposed by Kalman in 1960 [23], and a good introduction to the dynamic linear model and associated linear estimation techniques can be found in [41]. While the estimation procedure is best known as the Kalman Filter, we usually refer to the Dynamic Linear Model (DLM) itself, since this emphasizes both the linearity of the model and its dynamic nature, its ability to adapt to a system that is changing over time.

Figure 6-8 shows roll-out times and forecasts from both the static linear model and the DLM. Both series of estimates follow the same overall pattern, for both are linear functions of the same set of factors listed in Table 6.1. The DLM, however, adjusts its parameters to fit local conditions. Of course, these adjustments may be beneficial for some flights and not for others. Between 18:30 and 19:00 the forecasts generated by the DLM closely follow the observed values, improving significantly over those of the SLM. However, the SLM is more accurate between 18:00 and 18:30. Overall, the responsiveness of the filter improves performance: the RMSE drops from 7.8 min for the static model to 6.9 for the DLM. As was the case for the other forecasting procedures, the DLM did not succesfully forecast the

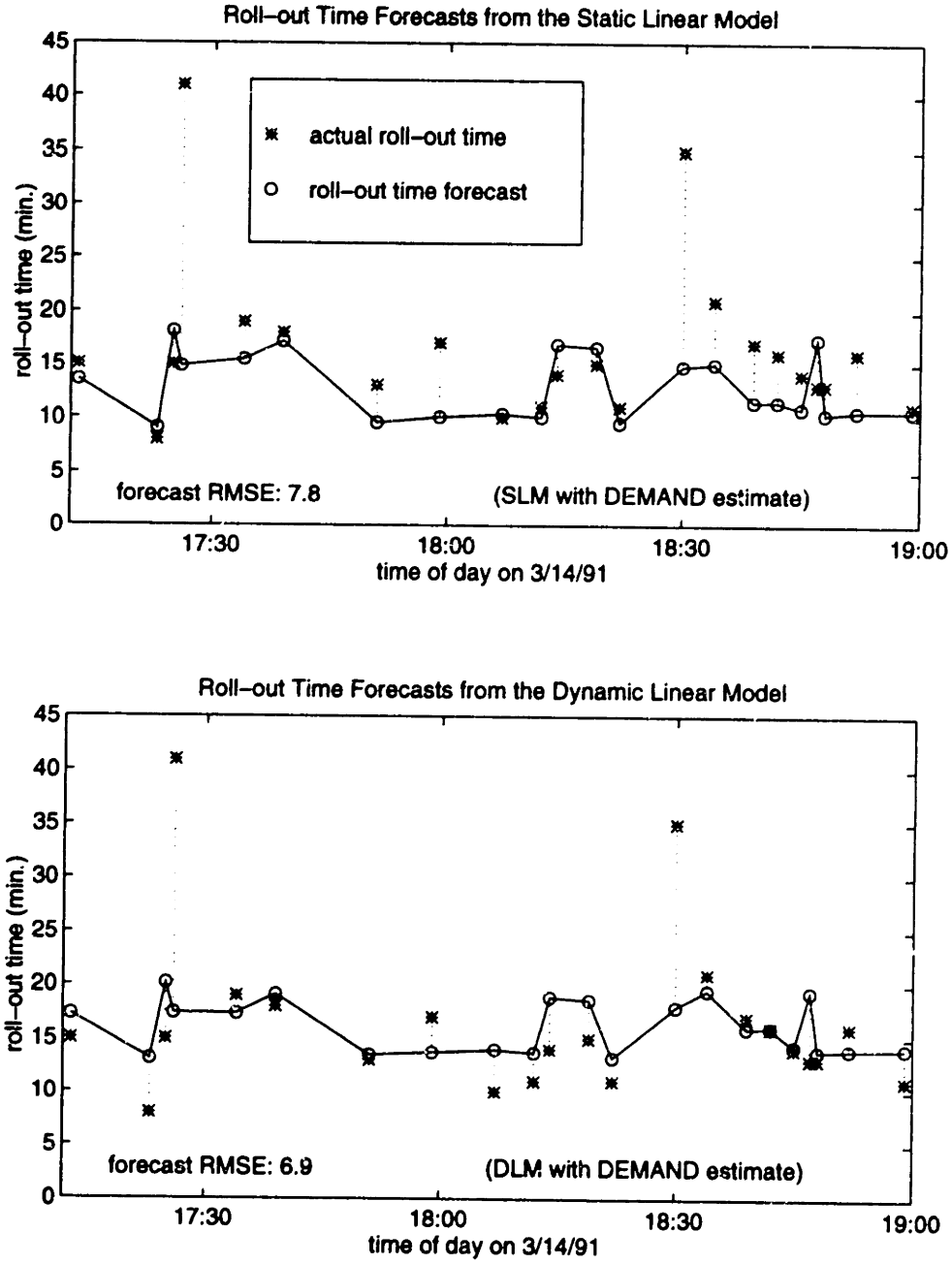


Figure 6-8: Actual roll-out times (*) and roll-out time forecasts (o) from the static linear model and the dynamic linear model. Both linear models use the DEMAND estimates.

extreme roll-out times shown in the figure.

While the DLM is a forecasting model with much flexibility, it does introduce more complexity in the form of parameters $\mathbf{W}$ and V. The observational variance V may be estimated from the residuals of the static regression model described in the last section. The system covariance specifies the volatility of the model coefficients. For the example presented in Figure 6-8, as well as for the empirical experiments presented in section 6.3, we assume that $\mathbf{W}$ is diagonal, so that the movement of one coefficient is uncorrelated with the movement of another. Let W_{ii} be a value on the diagonal of the matrix. Assume that all W_{ii} fall into one of two groups:

1. *Volatile* coefficients whose values change markedly over time and therefore have relatively high system variances. For these, we set $W_{ii} > 0$.

2. *Static* coefficients whose values do not change over time and therefore have no system variance. For these, we set $W_{ii} = 0$.

Intuition says that coefficients associated with service rates will be volatile, while those associated with the geography of the airport will be static. For example, the coefficient associated with the distance from USAir gates to particular runway configurations should change little over time, while the coefficients associated with service rates are affected by weather conditions or controller preferences. We would expect these to be more volatile.

Since conditions at a destination airport may change rapidly, coefficients for particular destinations (EWR, LGA and DFW) were also assumed to be volatile. Figure 6-9 shows the coefficient associated with flights to LaGuardia on August 25, 1991. The sudden evening increase in the coefficient may signal a traffic management problem such as weather difficulties in the New York area. The ability of the DLM to detect and respond to such changes may lead to superior forecast accuracy. This hypothesis will be tested in Section 6.3.

6.2.4 Combining Forecasts

We have described a variety of methods for estimating the parameter vector $\boldsymbol{\Theta}_k$. Once an estimate $\hat{\boldsymbol{\Theta}}_k$ has been found, equation 6.2 supplies the l-step forecast. This is the prior mean for the roll-out time $k + l$.

It is possible that if forecasts from two or more methods do not perform well, then a superior forecast may be produced by combining forecasts. This is especially true if the

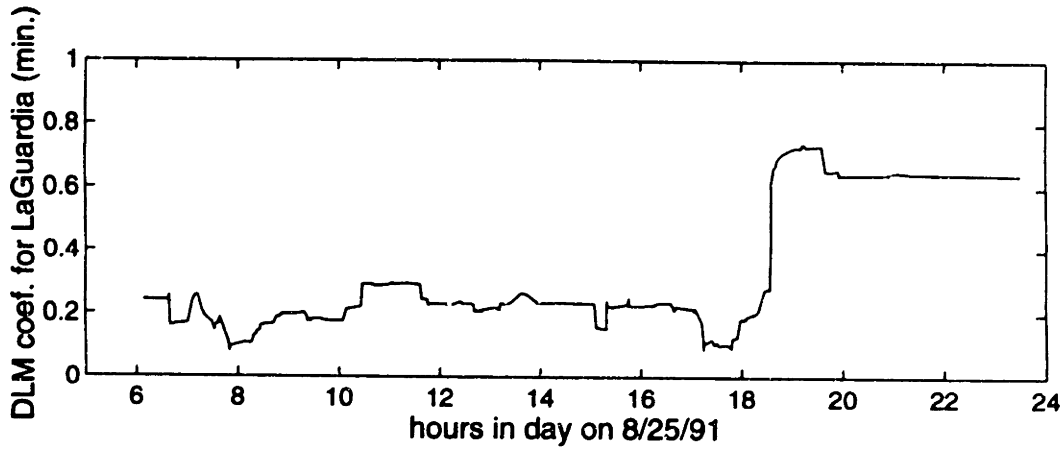


Figure 6-9: Changing value of the coefficient for flights to LaGuardia in the Dynamic Linear Model

forecasts complement each other, i.e. forecasts from some methods are most accurate when forecasts from the others are inaccurate. Given two unbiased forecasts $\hat{r}^a_{k+l}$ and $\hat{r}^b_{k+l}$, we may obtain a new forecast from the linear combination:

$$\hat{r}^c_{k+l} = \lambda \hat{r}^a_{k+l} + (1-\lambda)\hat{r}^b_{k+l} \tag{6.35}$$

Let σ_a^2 and σ_b^2 denote the variance of each forecast and let ρ denote the correlation coefficient between the forecast errors. It is not hard to show that the variance of the combination is minimized when:

$$\lambda = \frac{\sigma_b^2 - \rho\sigma_a\sigma_b}{\sigma_a^2 - \sigma_b^2 - 2\rho\sigma_a\sigma_b} \tag{6.36}$$

Note that when $\rho = 0$, $\lambda = \sigma_b^2/(\sigma_a^2 + \sigma_b^2)$. Intuitively, when the variance of the second forecast is relatively high, λ approaches one, and the first forecast receives more weight. For alternative methods for combining forecasts, see [6]. In the following empirical experiments we will test the forecast accuracy of each model as well as the accuracy of combinations of the most promising techniques.

The next section describes empirical experiments with each estimation procedure.

6.3 Empirical Tests of the Roll-out Time Models

In this section we compare the performances of the forecast algorithms. Comparisons are made between exponential smoothing, the lookup table, and the static and dynamic linear models. The empirical experiments were performed on the *testing data*, 2412 flights from Logan Airport during March and August, 1991. The testing data includes only those flights with known push-back times, 44% of all flights during the sample period. The tests demonstrate that while some procedures are superior for short horizons and others for longer horizons, all perform within a fairly narrow range. Results from exponentially weighted regression are not presented here; its performance was consistently inferior to all other regression procedures.

Before presenting the models and results, we must first settle a rather technical issue: what information is available to the forecasting model when the forecast is made? These models base their predictions on forecasts of environmental factors: the linear models use weather, runway and departure queue information to generate roll-out time predictions. In practice, these environmental forecasts must also be predicted. To simplify the analysis, we will use weather and runway observations instead of forecasts when generating roll-out time predictions. For example, if the ETMS were generating a prediction for a 9:00 AM roll-out at 8:00 AM, a forecast for the 9:00 AM weather state would be used. In these empirical tests, we will be using the *observed* 9:00 weather state and runway configuration to make our 8:00 roll-out time prediction. The models are using hindsight, and this may give them a slight advantage, for forecasts of weather do not always correspond with reality. However, the advantage should not be very great. Weather states often change slowly of time, and certain weather patterns, such as morning fog, are quite predictable. Tower controllers are often able to predict runway configurations with great accuracy. The stepwise regression analysis of section 6.1 also demonstrated that, given queue lengths and runway configuration, the direct influence of weather state on roll-out time is small. The models should not perform substantially worse if weather observations are replaced with weather forecasts in the models.

The models will not take advantage of hindsight when forecasts use N_k, the number of aircraft seen on the airfield when flight k pushes back. To simplify the analysis, the first set of tests below will use the observed values N_k when forecasts are made (results from

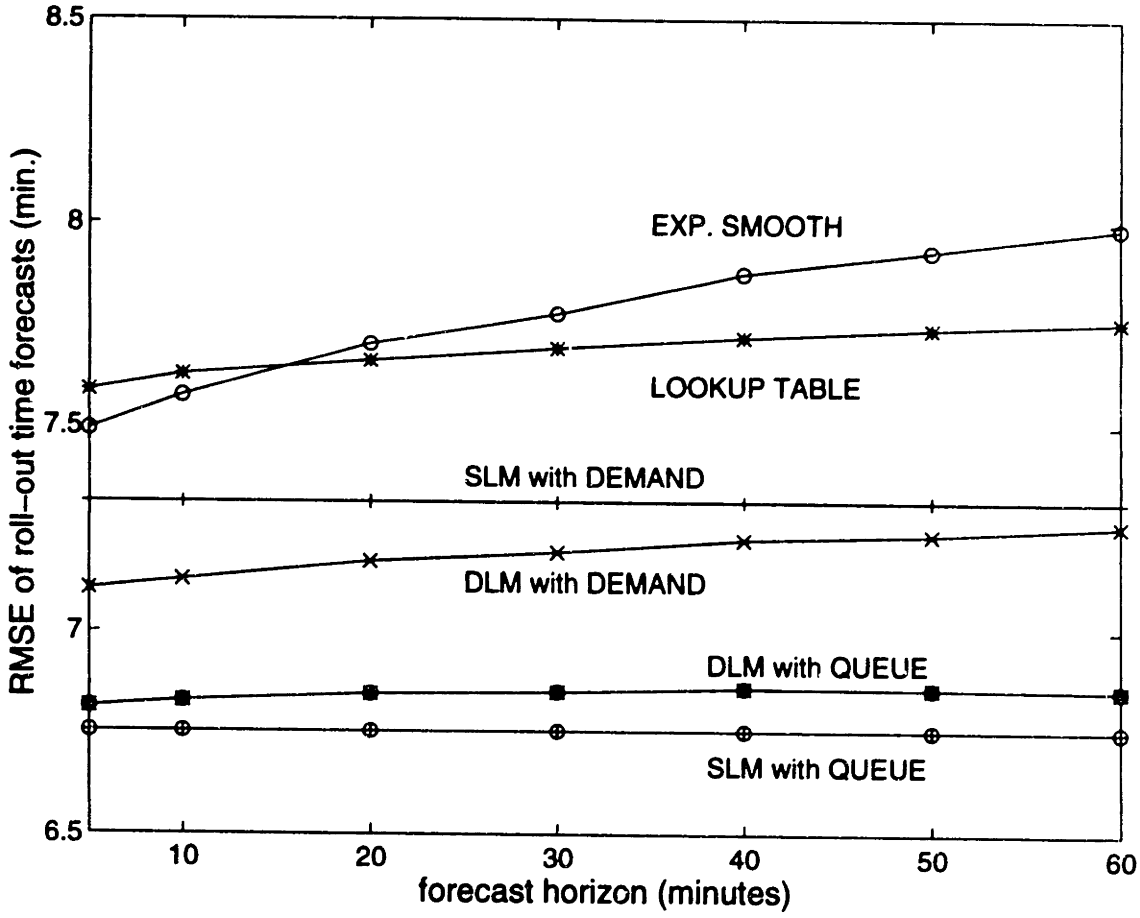


Figure 6-10: RMSE of roll-out time predictions for six forecasting models over horizons of 5 minutes to one hour. Only the 44% of flights with known roll-out times were used to generate the error statistics. Note that the y-axis has a range of only 2 minutes.

these tests are shown in Figure 6-10). With these results, we will determine the relative accuracy of the algorithms without the added variance injected by inaccurate forecasts of the congestion. The remaining empirical experiments, which will be conducted over forecast horizons as large as three hours, will use roll-out *predictions* $\hat{N}_k$ produced by the aggregate flow model. This will more closely approximate reality, for forecasts of N_k are not always accurate in actuality.

Following are details about the parameters used in each procedure:

Static Linear Models

Two static linear models were developed from the training data set. One includes departure congestion estimates N_k as a factor while the other replaces congestion estimates

with an estimate of departure demand in the 15-minute window around push-back. The departure congestion estimates will serve as a proxy for the departure queues seen by aircraft as they roll-out. In the following discussion, the two static models will be called "SLM with QUEUE" (static linear model with queue estimates) and "SLM with DEMAND" (static linear model with the simple DEMAND estimate). The factors included in the models are listed in Tables 6.2 and 6.1, respectively.

It should be emphasized that while it is called 'static,' the SLM does distinguish between good and bad weather. Terms in the SLM add or subtract roll-out time according to the individual and environmental characteristics of the flight.

Exponential Smoothing and the Lookup Table

For exponential smoothing (EXP), a 'fade factor' (f) of 0.9 produced the most accurate short-term forecasts. This is also the value of f used in the lookup table (LOOK). Our lookup table is categorized into three factors that are also used by the ETMS itself: weather category (VFR, MVFR, IFR and LIFR), time of day (morning, afternoon, evening) and length of flight (0-1 hr., 1-2 hrs., 2-3 hrs., 3+ hrs.). While this lookup table is similar to the present ETMS procedure for predicting ground transit time, direct comparisons with current ETMS performance are not possible since the ETMS does not yet receive push-back times and cannot separate gate delay from roll-out time.

Dynamic Linear Model

We implemented two dynamic linear models with the same factors as those used in SLM with QUEUE and SLM with DEMAND (see tables 6.2 and 6.1). These models will be referred to as "DLM with QUEUE" and "DLM with DEMAND", respectively. Parameter values for V and $\mathbf{W}$ also had to be specified. The observational variance V was estimated from the residuals of the static regression of the last section; we set $V = 20$ min. As described above, we divided the system variance coefficients W_{ii} into two categories: volatile and static. The model intercept and the coefficient for departure queue length were volatile and the remaining coefficients were static. For volatile coefficients, $W_{ii} = 0.05$, while $W_{ii} = 0$ for static coefficients. However, the performance of the DLM was not sensitive to the values of W_{ii}. Changing the variance matrix by a full order of magnitude had little effect on forecast

accuracy.

Prior to the application of the forecasting procedures, the 2412 flights in the testing data were ordered by take-off time. Exponential smoothing, the lookup table, and the static and dynamic linear models were all employed to make forecasts. Figure 6-10 displays the root mean squared error (RMSE) for forecasts produced by these procedures over forecast horizons ranging from five minutes to one hour. In this context, the forecast horizon is the time between forecast generation and actual push-back of the flight.

Perhaps the most important aspect of Figure 6-10 is the range of the y-axis: only two minutes. All four procedures have RMSEs that are within a small interval between 6.5 and 8.5 minutes. However, there are distinctions between procedures, and differences vary as the forecast horizon changes. Exponential smoothing is the most sensitive to the forecast horizon and its RMSE increases by 0.8 minutes between the 5-minute and one-hour forecasts. The lookup table and dynamic linear models also see a slight degradation in forecast accuracy as the horizon grows longer.

In general, the linear models produce more accurate forecasts than either the exponential smoother or the lookup table. The linear models which include airfield congestion estimates N_k (SLM and DLM "with QUEUE") outperform those with DEMAND estimates. This is not surprising, given the explanatory power of N_k demonstrated by the stepwise regression procedure in Section 6.1.

Given the examples of the previous section, we would also expect the DLM to outperform the SLM over short forecast horizons. The DLM, after all, is always responding to changes in observed roll-out times, while the static model was 'fitted' using historical data that did not overlap with the testing data set. This effect in seen in the models which use the simple DEMAND estimate (the middle two lines in Figure 6-10). With a forecast horizon of 5 minutes, the RMSE of the DLM is 0.8 minutes smaller than that of the SLM. The difference between the two shrinks as the forecast horizon grows. However, the positions of the SLM and DLM are reversed when congestion estimates N_k are used (see the bottom two lines of Figure 6-10). One explanation is that the congestion estimate N_k incorporates information about temporary changes in conditions. A decrease in the departure rate, for example, is accompanied by an increase in the departure queue. Since information about temporary conditions is already incorporated into the model in the form of N_k, the 'adaptable' DLM

coefficients in this model may be adjusting to purely random variations in roll-out times. These adjustments hinder, rather than help, forecast accuracy. Additional experiments found that increasing the responsiveness of the coefficients of the DLM decreased DLM performance, while decreasing responsiveness led to forecast accuracy as good as, but no better than, the accuracy of the static linear model.

In the experiments which generated Figure 6-10, the linear models had the benefit of hindsight when they used the estimates N_k instead of N_k forecasts from the flow models. When the N_k are replaced by departure queue size forecasts, $\hat{N}_k$, roll-out forecast accuracy suffers. In Figure 6-11 the top plot shows the RMSE for roll-out time forecasts from three models: SLM with QUEUE, SLM with DEMAND, and the lookup table. Three forecast horizons are shown: 5 minutes, 30 minutes and three hours. The bottom plot shows the mean absolute forecast errors for these procedures and forecast horizons. As was seen in Figure 6-10, all forecast procedures fall within a narrow band of about two minutes. This graph also demonstrates the effects of errors in N_k forecasts. While SLM with QUEUE is superior over the shorter forecast horizons, its performance is inferior to the others over the longest horizon. Errors in congestion predictions $\hat{N}_k$ lead to further errors in roll-out time predictions.

Table 6.3 shows results when observations are separated into good and bad weather periods. The first table displays the RMSE over all flights in the sample. These data were used to generate the top histogram in 6-11. The second and third tables show the RMSEs under non-VFR and VFR weather, respectively. The middle (bad weather) table shows a slightly wider variance in forecast performance. Under poor conditions the DLM and static models have a one-minute advantage over the lookup table for short forecast horizons. Over the three-hour horizon, linear models using DEMAND are preferable. At the bottom of Table 6.3 we see that under good weather, the procedures are more tightly grouped: all are within one-half minute for the shortest and longest forecast horizons.

We also combined forecasts from each of these models according to the procedures described in section 6.2.4. However, none of the forecast combinations were superior to the models described above. The lack of improvement is explained by the high correlation in forecast errors between forecasting models. All models had forecast error correlations in a range from 0.8 to 0.95. Table 6.4 displays the pairwise correlation coefficients between models. In general, a large forecast error by one model was mirrored by large errors from

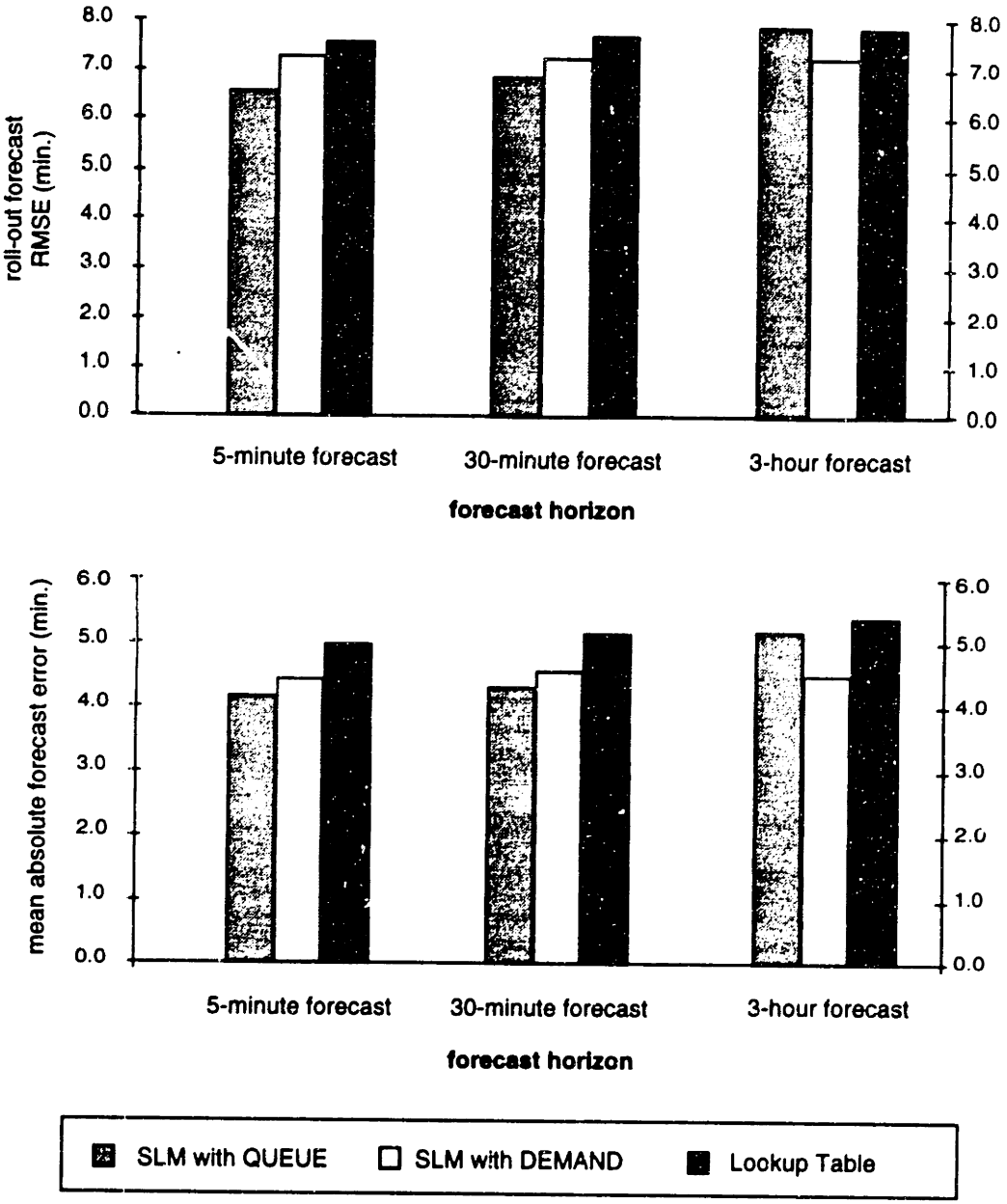


Figure 6-11: Square root of mean squared forecast error (top) and mean absolute forecast error (bottom) of roll-out time forecasting models over short and long forecast horizons.

All flights:

forecast horizon	SLM w/ QUEUE	DLM w/ QUEUE	SLM w/ DEMAND	DLM w/ DEMAND	LOOKUP
5 min.	6.8	6.8	7.3	7.1	7.6
30 min.	6.9	7.0	7.3	7.2	7.7
3 hours	7.9	7.9	7.3	7.3	7.9

Non-VFR flights:

forecast horizon	SLM w/ QUEUE	DLM w/ QUEUE	SLM w/ DEMAND	DLM w/ DEMAND	LOOKUP
5 min.	8.3	8.3	9.2	8.7	9.3
30 min.	8.6	8.6	9.4	8.9	9.6
3 hours	10.2	10.0	9.5	9.0	9.9

VFR flights:

forecast horizon	SLM w/ QUEUE	DLM w/ QUEUE	SLM w/ DEMAND	DLM w/ DEMAND	LOOKUP
5 min.	5.7	5.8	5.9	5.9	6.3
30 min.	6.0	6.1	6.0	6.1	6.5
3 hours	6.7	6.8	6.3	6.3	6.9

Table 6.3: Roll-out time forecast RMSE for all data (top), bad weather (middle) and good weather roll-outs (bottom).

model	SLM Q	SLM NQ	LOOK	EXP
DLM	0.91	0.80	0.81	0.90
SLM Q	-	0.93	0.85	0.84
SLM NQ	-	-	0.86	0.80
LOOK	-	-	-	0.89

Table 6.4: Correlation coefficients of forecast errors between models.

forcast horizon	forecasting model	MAE	% decrease	RMSE	% decrease
	mean	5.4	-	7.8	-
5 min.	LOOKUP	5.0	7%	7.6	3%
	SLM with QUEUE	4.1	24%	6.8	12%
	SLM with DEMAND	4.5	17%	7.3	6%
3 hours	LOOKUP	5.3	2%	7.9	-1%
	SLM with QUEUE	5.1	6%	7.9	-1%
	SLM with DEMAND	4.5	17%	7.3	6%

Table 6.5: Performance of roll-out time forecasting models. Mean absolute forecast error (MAE) and root-mean squared forecast error (RMSE) of roll-out times from an aggregate mean and from a roll-out time model. A negative change in RMSE percentage indicates that the model produced forecasts with larger RMSE than the overall mean.

the others.

6.4 Conclusions

We began this chapter by developing two static linear models of roll-out time. Using historical data, we were able to detect significant factors which produce variations in roll-out time. These include geographical factors (distance from a carriers' gates to runways), bad weather, and the duration of a flight. The factor with the greatest impact is the amount of congestion on the airfield, represented in our model by N_k, the number of aircraft rolling out.

These models produced small gains in forecast accuracy which are summarized in Table 6.5. The table compares the two linear models and the lookup table with the simplest forecast possible: a constant forecast of 18 minutes, the mean roll-out time in the sample. Over a forecast horizon of 5 minutes, the static linear model which includes N_k (SLM with

QUEUE) improves mean absolute error (MAE) by 24% and RMSE by 12%. For a forecast horizon of three hours, however, this model is little better than the overall mean while an alternate model, SLM with DEMAND, produces more accurate forecasts. SLM with DEMAND replaces N_k with a simple estimate of departure demand during the period surrounding each roll-out. The lookup table is less accurate than both linear models over the the shortest forecast horizon, and is only as good as the mean over the three-hour horizon.

We attempted to improve the static linear models by adding a dynamic component. One promising approach was to develop a dynamic linear model (DLM) which was similar in form to the static models but has coefficients which respond to temporary changes in roll-out times. The DLMs did perform as expected, and occasionally outperformed the static linear models. However, for short-term forecasts, the SLM with N_k was as successful as any dynamic model, for N_k itself is affected by temporary changes in conditions on the airfield. For long-term forecasting, the adaptability of the DLM is unwelcome, and the static model SLM with DEMAND is as reliable. Given these results, we do not recommend the adoption of a DLM for this application.

The relatively small improvements in forecast accuracy provided by the most sophisticated models and the failure of forecast combinations to improve accuracy highlight the fact that all proposed forecast algorithms were unsuccessful in predicting the largest roll-out times, those on the 'right-hand tail' of the distribution. The two outlying Pan Am flights in Figure 6-8 were not unusual. Roll-out times of similar length may be found during almost any two-hour period. These large deviatiations can sometimes be explained in hindsight, but seldom can be anticipated given the available data.

As with the gate delay models, the carriers may provide information that will allow more accurate forecasts of the largest outliers. For example, American Airlines now has a "Company Arrival Control" (CAC) procedure when capacity is reduced at a major hub such as Chicago or Dallas/Ft. Worth [22]. Flights to the hub may be delayed on the ground by over an hour in order to reduce demand at the congested destination. Ground delays are assigned by American itself and may not be communicated to the FAA. According to gate personell at Logan Airport, many of the flights with the longest roll-out delays have pushed-back in order free the gate for arrivals. The flights then park in a waiting area on the apron until the assigned delay has expired. These delays would appear as remarkably high, unexplained roll-out times in the data, even though the precise length of the delay is

known far in advance by the airline. With improved communication, further improvements in forecast accuracy may be possible.

However, even with FAA/industry cooperation, it is likely that there will always be a significant amount of uncertainty associated with roll-out time predictions. Given this uncertainty, there may be opportunities to update forecasts after the initial forecasts have been generated by these models. Procedures for updating forecasts are presented in the next chapter.

Appendix 6A: Example of Information Loss in Exponentially Weighted Regression

We wish to estimate $\boldsymbol{\Theta} = (\theta_1, \theta_2)'$ after $N = m + n + 1$ observations. In particular, we will find the limiting estimates, $\hat{\theta}_N = (\hat{\theta}_{N,1}, \hat{\theta}_{N,2})'$, as $n \to \infty$.

The solution for $\hat{\theta}_N$ is given in equation 6.26, with $\underline{\mathbf{F}}_N$ defined in equation 6.20. By working through the matrix multiplication we find that

$$\begin{pmatrix} \hat{\theta}_{N,1} \\ \hat{\theta}_{N,2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} f^{N-i} & \sum_{i=1}^{m} f^{N-i} + 1 \\ \sum_{i=1}^{m} f^{N-i} + 1 & \sum_{i=1}^{m} f^{N-i} + 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} f^{N-i} r_i \\ \sum_{i=1}^{m} f^{N-i} r_i + r_N \end{pmatrix} \tag{6.37}$$

The calculations simplify as the period of good weather increases in length and $n \to \infty$. If $f < 1$, $\lim_{n\to\infty} \sum_{i=1}^{m} (f^{N-i}) + 1 = \lim_{n\to\infty} f^n \sum_{i=1}^{m} (f^{m+1-i}) + 1 = 1$. Since the product of two absolutely converging series converges to the product of the series, as n grows large,

$$\begin{pmatrix} \hat{\theta}_{N,1} \\ \hat{\theta}_{N,2} \end{pmatrix} \approx \begin{pmatrix} (1-f)^{-1} & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} f^{N-i} r_i \\ r_N \end{pmatrix} \tag{6.38}$$

$$= \begin{pmatrix} (1-f) \sum_{i=1}^{N-1} f^{(N-1)-i} r_i \\ r_N - \left[(1-f) \sum_{i=1}^{N-1} f^{(N-1)-i} r_i \right] \end{pmatrix} \tag{6.39}$$

$$= \begin{pmatrix} \hat{\theta}_{N-1,1} \\ r_N - \hat{\theta}_{N-1,1} \end{pmatrix} \tag{6.40}$$

These values for $\hat{\theta}_{N,1}$ and $\hat{\theta}_{N,2}$ have a straightforward interpretation. The estimate $\hat{\theta}_{N,1}$ is equal to the exponentially weighted average of all observations before the Nth observation (see equation 6.13). The estimate $\hat{\theta}_{N,2} = r_N - \hat{\theta}_{N,1}$ is the residual in the Nth observation which is not explained by the intercept, $\hat{\theta}_{N,1}$. The estimate $\hat{\theta}_{m,2}$ which was derived after the first m observations has been discarded, and the new estimate is based on a single

observation, r_N.

Chapter 7

Optimal Updating of Forecasts

The gate delay and roll-out models developed in previous chapters improve take-off time forecast accuracy, but do not eliminate the uncertainty associated with take-off times. These models make predictions based on the behavior of previous flights that operated under similar conditions. Since the ETMS gathers information in real-time, there may be opportunities to improve the forecasts with more recent information. For example, if the ETMS reports that a flight has pushed back, or that a flight has reached the head of the runway, the forecast might be updated.

Unfortunately, the ETMS does not yet report ground flight events, such as push-backs or movements on the taxi-ways. The aircraft is invisible to the system while it is on the ground. An active flight first appears in the system when it takes off and a DZ message is generated. However, even the fact that a take-off has not occurred may be enough information to improve a forecast. If thirty minutes have passed since a scheduled push-back and the aircraft has not yet taken off, a new forecast would reflect the information that the ground transit time (GTT) is at least thirty minutes.

Information about each flight arrives frequently to the ETMS. If the only objective were forecast accuracy, forecasts of take-off time should be updated as soon as any new information about GTT arrives. However, forecast updates may be expensive, requiring data collection and computer processing. They can also be disruptive, diverting attention away from other tasks. This section investigates methods for balancing update costs with forecast accuracy when predicting the take-off time of an aircraft. However, the techniques developed in this section are flexible and may be applied to forecasts of other events, such

as the arrival of a thunderstorm or the release of an item from a job shop.

In the traditional forecasting problem, the optimal predicted time of an event minimizes a single, expected cost, such as the squared deviation between the forecast and the actual time of the event. This criterion implies that the cost is evaluated once, possibly when the event occurs and the accuracy of the forecast can be determined. In practice, the consequences of an inaccurate forecast may build up over time and the significance of a forecast error may depend on its timing relative to the event. For example, air traffic management decisions based on take-off time forecasts may be made far in advance or immediately before the actual take-off. In addition, the traditional definition of forecast optimality evaluates a forecast in isolation, under the assumption that it will not be revised. However, as was noted above, the optimal forecasting procedure may not be a single prediction but may be a *sequence* of forecasts which is produced as information about the event becomes available. Such a forecasting procedure must include both the times at which the forecasts are produced (an *update schedule*) as well as the forecasts themselves.

A measure of the utility of a sequence of forecasts must be based on the accuracy of all the forecasts made at various times before the event of interest. Therefore, we propose a cost function which evaluates a sequence of forecasts rather than a single prediction. A time-dependent weighting function allows forecast errors to have more or less influence, according to the proximity of the forecast to the predicted take-off time. In order to prevent an excessive number of forecast revisions, a cost is also assessed on each update that occurs before the event. Optimizing over this objective function leads to update schedules and forecasts which produce accurate information when it is needed, while simultaneously holding down the expected number of updates.

The production of forecasts has been a subject of study throughout the physical and social sciences. This paper adopts the notation used by Seidmann and Smith [32], who investigated optimal predictions of completion time for jobs in a production system. Seidmann and Smith's paper treats each forecast as an isolated prediction and does not consider the design of a sequence of forecasts.

The notion of an update schedule is similar to the inspection schedules developed in numerous articles on reliability theory. These articles discuss systems with random times until failure. Barlow, Hunter and Proschan designed inspection schedules which minimize a function of both inspection costs and the expected lapsed time between system failure and

inspection [5]. Later work produced schedules for systems which travel through three states: healthy, failed but asymptomatic, and failed with symptoms (see, for example, [34], [28], and [42]). These models were used to determine examination schedules for chronic diseases as well as for machine failures. The cost function defined in these articles is significantly different from the function proposed here. The inspection schedules derived in these articles determine whether the system is healthy. In terms of reliability theory, the schedules derived here both check on the health of the system and produce forecasts of the remaining time until failure.

The remainder of this chapter is organized as follows. Section 7.1 links the ground transit time models of the previous chapters with the update procedures developed in this chapter. The update procedures will be driven by prior probability distributions of take-off times. Our models of gate departure delays and roll-out times generate point forecasts with some distribution of errors around the forecasts, and these errors may be used to estimate a distribution for take-off times. As an example, Section 7.1 demonstrates how we may use the residuals from a roll-out time model to estimate a probability distribution for roll-out time predictions. Section 7.2 describes the updating procedure currently used by the ETMS and provides alternate procedures. Section 7.3 introduces a cost function which includes penalties for both forecast errors accumulated over time and update costs. In section 7.4 a procedure is described for finding optimal forecasts, given an update schedule. In addition, a dynamic program finds an optimal update schedule, given a sequence of forecasts. A heuristic combines these results to find a near-optimal combination of update schedules and forecasts. In section 7.5 the procedures are tested in numerical examples which approximate situations confronting an air traffic manager. The procedures developed here are compared to the procedure currently used by the ETMS. The results demonstrate that when prior uncertainties are large, the optimal procedures developed in this chapter have higher forecast accuracy and lower update costs than the current system. For moderate levels of uncertainty, simple heuristics provide comparable benefits.

7.1 Generating Probability Distributions from Roll-out Models

The updating procedures that will be described in this chapter generate forecasts from a prior probability distribution of take-off times. However, the forecasting procedures described so far have been used to generate *point* forecasts rather than probability distributions. This section provides a bridge between the roll-out models of the previous chapter and the update procedures developed in this one. In this section we fit a probability distribution to the residuals from the static linear roll-out model (SLM NQ). Similar procedures could be used for other roll-out models and the gate departure delay models. Since gate departure delays and roll-out times are (approximately) statistically independent, a probability density function for ground transit times can be generated by simple convolution of the gate departure delay and roll-out time distributions.

An important question is whether a single pdf will suffice for all roll-out time forecasts. The answer, clearly, is 'no,' for even the point forecast from the static linear model,

$$\hat{r}_k = \mathbf{F}_k' \hat{\boldsymbol{\Theta}} \tag{7.1}$$

varies with the factors $\mathbf{F}_k$. In addition, the variance of roll-out times does vary slightly with the underlying conditions: the data exhibit *heteroscedasticity*. Roll-outs during bad weather, for example, tend to have a larger variance than good-weather flights. Therefore, one might collect empirical distributions for flights under certain weather conditions, runway configurations, etc. However, as was the case for the lookup table, this procedure would become cumbersome if a large number of factors were used. Instead, we assume that each roll-out distribution has the same shape, but that each flight's distribution has a customized location based on the point forecast $\hat{r}_k$. For reasons that will become clear, we also assume that roll-outs follow a gamma distribution. Parameters for the gamma distribution will be estimated from the residuals of the static linear roll-out model.

Even in the absence of heteroscedasticity, the assumption of a constant roll-out distribution is an approximation. If we assume that the static linear model defined by equation (6.17) is true, then sources of error in $\hat{r}_k$ can be divided into two parts: errors in the specification of $\hat{\boldsymbol{\Theta}}_k$ and random errors added to each observation. An estimate of the vari-

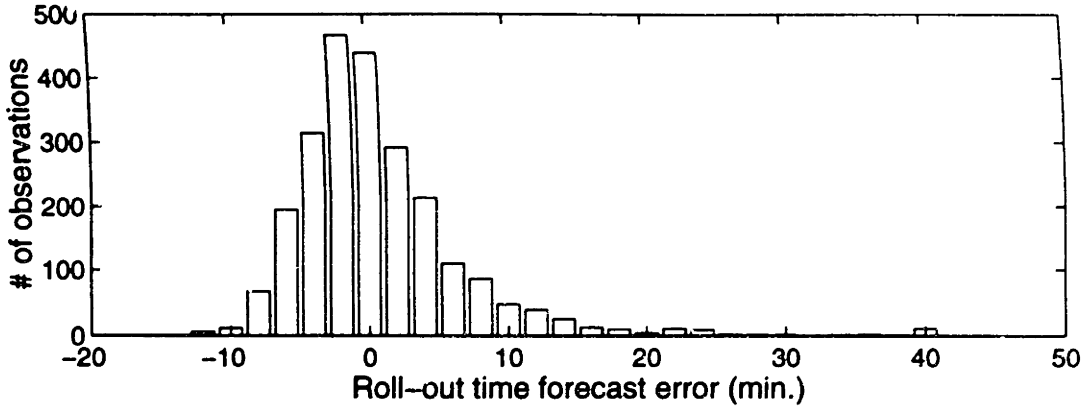


Figure 7-1: Histogram of residuals from the static linear model (SLM NQ). There are 11 residuals larger than 50 minutes.

ance of these random errors, s^2, can be calculated from the residuals of the model according to equation 6.5. Let $\underline{\mathbf{F}}$ be the matrix of factors used in the training data. An estimate of the variance of the forecast itself is:

$$\text{Var}(\hat{r}_k) = s^2 \left(1 + \mathbf{F}_k'(\underline{\mathbf{F}}\underline{\mathbf{F}}')^{-1}\mathbf{F}_k\right) \tag{7.2}$$

This variance is not constant over all flights, for the second term in the parentheses of the right-hand side represents the proportion of variance contributed by the errors in $\hat{\boldsymbol{\Theta}}$ which multiply the factors $\mathbf{F}_k$. However, there were over 2400 flights represented in $\underline{\mathbf{F}}$, and the coefficient values $\hat{\boldsymbol{\Theta}}$ have relatively small standard errors. As a result, the term $\mathbf{F}_k'(\underline{\mathbf{F}}\underline{\mathbf{F}}')^{-1}\mathbf{F}_k$ varies between 0.002 and 0.05 in our sample; it is negligible for all flights. We can assume that $\text{Var}(\hat{r}_k) \approx s^2$ for all k. Heteroscedasticity, or variation in the underlying noise estimated by s^2, is of greater concern. It may be necessary to determine an efficient partition of flights, grouping together flights with similar distributions.

However, for this initial experiment we will assume that all flights are adaquately described by the same probability distribution. The distribution will be shifted so that it is centered on $\hat{r}_k$. Figure 7-1 shows the distribution of forecast errors e_k from the static linear model. The estimated variance of the noise from this sample is $s^2 = 62$ min.^{2} Like the roll-out times themselves, these residuals have a 'long right tail.' Let $f_e(e_0)$ be the probability density function which describes these model errors. Given roll-out prediction $\hat{r}$, the pdf for the roll-out is $f_r(r_0) = f_e(r_0 - \hat{r})$. The error distribution is shifted so that 'zero error' corresponds to a roll-out equal to $\hat{r}$.

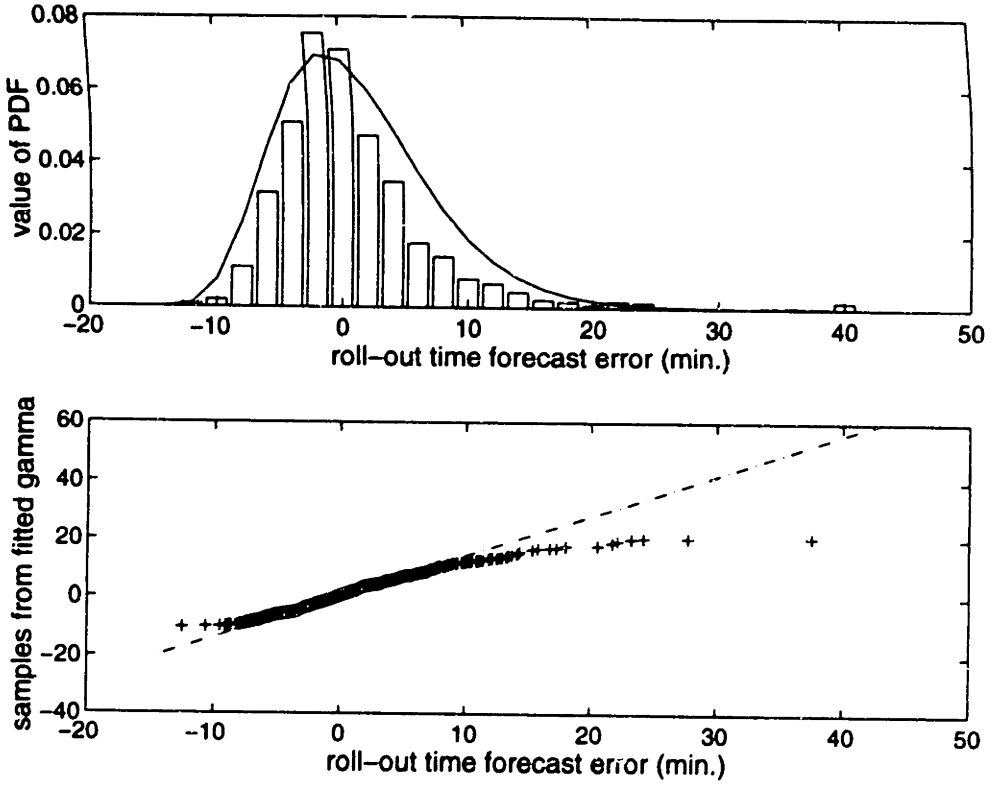


Figure 7-2: Gamma distribution fitted to forecast error histogram, with the forecast errors superimposed (top). The bottom plot shows a quantile-quantile plot of forecast residuals and the fitted gamma distribution.

To estimate the distribution $f_e(e_0)$ from Figure 7-1, we might consider a simple empirical distribution. However, it will be computationally more efficient to fit a parametric distribution to the data. In the following sections we will be able to derive algebraic expressions for optimal forecasts given prior normal or gamma distributions for take-off times. From the shape of Figure 7-1, and since roll-out times are non-negative, a shifted gamma distribution is a likely candidate for $f_e(e_0)$.

Let $f_\gamma(x_0; a, b)$ be a gamma distribution with parameters a and b. We will assume roll-out times are distributed as:

$$f_r(r_0) = f_e(r_0 - \hat{r}) \tag{7.3}$$

$$= f_\gamma(r_0 - \hat{r} - c; a, b) \tag{7.4}$$

Roll-out times are distributed as a gamma distribution, shifted twice: once by c min. so that the gamma distribution corresponds to the error distribution in Figure 7-1, and once so that the roll-out distribution is centered on $\hat{r}$. To complete the specification of $f_r(r_0)$,

we must estimate parameters a, b, and c from the data.

Since the gamma distribution is nonzero for non-negative values, we set $c = \min(e_k) + \delta = -14 + \delta$ min., where δ is some small nonnegative number. The addition of δ ensures that the lowest error value is nonzero after shifting. We set $\delta = 0.01$ min., although the values of the estimated parameters a and b were insensitive to the exact value of δ as long as δ was between 0 and 1.

The estimates for a and b may be derived using a method of moments or a maximum likelihood procedure [9]. Both were employed, and the maximum likelihood procedure produced a distribution which more closely fit the histogram in 7-1. The estimated parameters were: $\hat{a} = 6.5$ and $\hat{b} = 2.4$. The mean and variance of the fitted distribution are $\hat{\mu} = \hat{a}\hat{b} = 15.6$ and $\hat{\sigma}^2 = \hat{a}\hat{b}^2 = 37.4$ min.^{2} The mean of the fitted distribution matches the mean of the empirical distribution, but the fitted variance is significantly lower than the observed variance, 62 min.^{2}

The fitted gamma distribution is shown in Figure 7-2 with a superimposed histogram of model errors. A comparison of the forecast errors and the fitted gamma distribution is shown in the quantile-quantile plot in the bottom of the figure. As was true in Figure 6-4 in the last chapter, there are large, unexplained residuals in the data, and these produce the curved tail in the quantile-quantile plot, as well as the difference in variance between the observed and fitted distributions. Model errors are not quite distributed as a gamma; a distribution with a longer right tail would produce a better fit. However, the gamma distribution produces a good fit for the large majority of observations. A gamma distribution with the parameters derived above will serve as one example in the empirical results of Section 7.5.

7.2 The Aircraft Departure Update Problem

Every five minutes, ETMS take-off time forecasts are distributed to air traffic managers throughout the country, who use the forecasts to anticipate congestion and make local flow control decisions. Let random variable g represent an aircraft's ground transit time, the difference between the actual take-off time and the scheduled departure time. Before the scheduled departure time of each aircraft, the gate departure delay and roll-out models will generate a forecast for g. Later, information about the flight's status might be used to update the forecast.

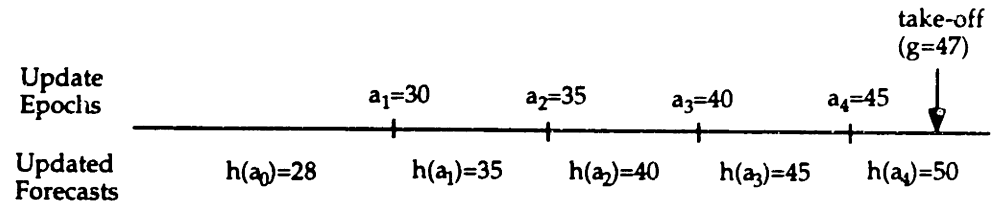


Figure 7-3: Constant interval updating with an original forecast of 28 minutes

When a flight takes off, the ETMS is notified almost immediately. Conversely, the *absence* of a take-off signal represents information (' the flight has not yet taken off') that might be used by the ETMS. At any time, the ETMS has the opportunity to update the take-off time forecast based on this information. In fact, each flight has a pre-defined *update schedule*, a schedule of times when forecasts are updated, given that a take-off has not been observed. The update schedule is followed until actual take-off occurs. Define a forecast update at time a_k to be an *update epoch* and the ordered sequence of update epochs $\mathbf{a} = \{a_0, a_1, \ldots, a_n\}$ is an update schedule. The update epoch a_0 is defined to be $-\infty$ and represents the production of the initial forecast of g. This forecast will be updated at a_1 as long as a take-off has not yet been observed (i.e. if $g > a_1$). A final update is performed whenever a take-off does occur.

Associated with each update schedule is an ordered sequence of updated forecasts, $\mathbf{h(a)}$. The forecast $h(a_i)$ is the forecast of g which is in effect from a_i until the next update at a_{i+1}, so that $\mathbf{h(a)} = \{h(a_0), h(a_1), \ldots, h(a_n)\}$. The following procedure is used by the ETMS to produce a sequence of forecasts.

Constant Interval Updating: In the current implementation of the ETMS, forecasts of flight departure times may be updated every five minutes. Five minutes is the cycle time of the ETMS, the interval between epochs when the ETMS refreshes the video screens of the air traffic managers. After making an initial prediction, $h(a_0)$, the system does not update the forecast until the beginning of the first five-minute cycle *after* $h(a_0)$. Thereafter it repeatedly adds five minutes to the forecast until the flight departs. Therefore, update epoch $a_1 = \lceil h(a_0) \rceil^5$, where $\lceil x \rceil^5$ is the beginning of the first five-minute interval after x. The updated forecast $h(a_1) = a_1 + 5$ and the next update epoch $a_2 = h(a_1)$. Similarly, $h(a_2) = a_2 + 5, a_3 = h(a_2)$, etc. If there is no observed take-off by one hour, the flight is marked as canceled. If the flight eventually does take-off, it is immediately re-activated but is seen as a surprise arrival to the users of the system.

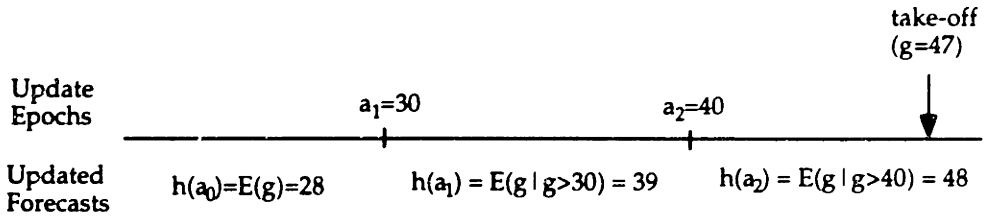


Figure 7-4: Discrete conditional updating with an original forecast of 28 minutes. Update epochs occur only at the beginning of a five-minute cycle.

Figure 7-3 illustrates an example of this update procedure. The initial forecast $h(a_0) =$ 28 min. The beginning of the first cycle after 28 min. is 30 min. An update occurs at 30 min., producing a new forecast of 35 min. The update at 35 min. produces a new forecast of 40 min., and this continues until the flight departs at 47 min. There is a total of six updates in this example: our initial forecast $h(a_0)$, updates at 30, 35, 40, and 45 minutes, and a final update when the flight does depart.

The number of *scheduled* updates is fixed: one every five minutes until an hour passes. However, the number of updates that are actually performed is a random variable with a minimum of two: one 'update' for the initial forecast and one at the true time of take-off. Note that the final update at take-off occurs instantly and is not restricted to be at any scheduled update epoch a_k, since a departing aircraft immediately appears on the radar screens of the local air traffic managers and they need not wait for an update from the ETMS.

The constant interval procedure has two fundamental flaws. The first is that the updated forecasts *ignore distributional information about g.* While it is simple to add five minutes to each forecast, there is no reason to believe that such a rule minimizes a relevant measure of expected forecast error. The second flaw echoes the first. The timing of the update epochs is unlikely to be optimal, and the ETMS may schedule too many updates. This can have serious consequences for the overall system. Each update requires significant computer processing to alter the large data structures that keep track of flights. In addition, frequent updates can confuse air traffic managers and can lead to distrust in a system that cannot 'make up its mind.' Thus, what little is gained in accuracy can be lost in greater expense and frustration.

Now consider two alternative procedures for generating update schedules and forecast sequences of aircraft departure times. The first is driven by the conditional means of the

distribution of g. The second is a continuous updating scheme which should outperform the discrete procedures in terms of forecast accuracy.

Discrete Conditional Updating: The update epochs are determined in the same manner as the constant interval epochs. Once a forecast has passed, the next update is scheduled at the beginning of the following cycle. However, each updated forecast $h(a_k)$ is the *conditional expectation* of g given that $g > a_k$. As is well-known in decision analysis, the conditional expected value minimizes the mean squared error of the forecast. Other forecasts that can be derived from the conditional distribution, such as the median, would minimize other loss functions.

As in our previous numerical example, the initial forecast is 28 min. and the next update epoch a_1 occurs at the beginning of the following cycle, at 30 min. (see Figure 7-4). Suppose at time 30 we determine that $E(g|g > 30) = 39$ min., so that the first updated forecast is $h(a_1) = 39$ min. The next update occurs at 40 min., the beginning of the following cycle. Finally suppose that the second conditional forecast, $h(a_2)$, is 48 minutes, and the flight departs before the next update can occur.

In this example, there were four updates. We will see that when compared with constant interval updating, discrete conditional updating reduces the number of expected updates while producing forecasts that reduce the expected forecast error.

Continuous Conditional Updating: Suppose that the ETMS has an infinitesimal cycle time, so that forecast updates may be released in an essentially continuous manner. Such a system would allow continuous conditional updating, so that $h(t) = E(g|g > t)$ for all t. Since the system would always produce the most accurate forecast possible, it would outperform its discrete counterparts but would require an essentially infinite number of updates for each flight. For certain cost fuctions, such a system establishes a *lower bound* on forecast error.

As these examples suggest, there is a tradeoff between accuracy and update frequency. To measure the utility of alternate update procedures, the next section defines the costs associated with any particular update schedule. Later sections derive schedules and forecasts that minimize these costs. Throughout, the narrative will refer to the aircraft departure update problem described above, but the concepts are applicable to any situation in which forecasts may change over time.

7.3 Cost Function for a Sequence of Forecasts

This section describes a cost function for forecasts that are updated as the departure time approaches. The smallest 'building block' of the function is $C(g_0 - h(t))$, the cost of the difference between an actual departure time g_0 and $h(t)$, the forecast at time t. The function $C(g_0 - h(t))$ describes the cost of a *single* forecast error. However, we wish to assess the accuracy of an update schedule which specifies a number of forecasts over a number of time periods. The cost function proposed here accumulates forecast error over time while weighting these errors according to a time-varying function. For example, errors made far before the proposed departure time of an aircraft may be given less weight than those made close to the proposed departure, when a local traffic manager is most interested in the forecast. The total forecast error cost is found by integrating the weighted forecast errors up to the actual take-off time. In addition to the forecast error, a cost K_u is assessed for each forecast update. Given this cost formulation, an expression for the expected forecast and update cost is derived in terms of $f_g(g_0)$, the prior pdf of g.

Since the relative costs of updates and forecast accuracy may not be clear, it may be difficult to specify K_u. However, an analyst will be able to examine the trade-off between forecast accuracy and update cost by varying K_u and repeatedly solving for the optimal update schedule.

7.3.1 Cost Function for a Given Departure Time

Many well-established methods exist for assessing forecast accuracy and for designing optimal forecasts. For a single forecast we will adopt the general framework proposed by Seidmann and Smith [32], which allows the cost of an inaccurate forecast to assume a variety of functional forms.

Given a forecast $h(t)$ at time t and an actual departure time, g_0, specify a cost function C on the forecast error $g_0 - h(t)$. In order to include many possible functional forms, define two cost functions for each forecast. One cost is assessed if the flight departs before the forecast and another if the flight departs after the forecast:

$$C(g_0 - h(t)) = \begin{cases} C_r(h(t) - g_0) & \text{if } g_0 < h(t) \\ C_l(g_0 - h(t)) & \text{if } g_0 > h(t) \\ 0 & \text{if } g_0 = h(t) \end{cases} \tag{7.5}$$

Cost function C_r applies when there is an early departure before the predicted departure time while C_l applies when there is a late departure after the predicted time. Throughout this discussion, assume that the functions C_r and C_l are monotone increasing, strictly convex, twice differentiable, and vanish at the origin. This includes many reasonable cost functions and enables the derivation of simple expressions for the optimal forecasts. A numerical example which is presented later in this chapter uses a linear cost function which weights the two types of errors by constants K_r and K_l.

The cost function C is a static one, for it does not take into account the fact that we may update our forecast as time progresses. Therefore, define the forecast error cost for an actual departure at time g_0 to be:

$$\begin{aligned} C_F(g_0; \mathbf{a}, \mathbf{h}(\mathbf{a})) &= \int_{-\infty}^{g_0} w(t) C(g_0 - h(t)) dt \\ \text{where} \quad & w(t) \geq 0 \end{aligned} \tag{7.6}$$

The weighting function $w(t)$ reflects the relative importance of forecast error $g_0 - h(t)$ at time t. The forecast $h(t)$ is determined by the update schedule and may be revised during update epochs prior to g_0. This cost function weights the forecast error by $w(t)$ and accumulates this error until the departure of the flight. Besides the requirement that $w(t)$ be nonnegative, we will have no restrictions on its functional form. For example, in the numerical example given later, $w(t)$ is a step function which rises 1/2 hour before the prior expected take-off time. In general $w(t)$ may be used to penalize forecast errors that may be made when the forecasts are used to inform consequential decisions.

The cost $C_F(g_0; \mathbf{a}, \mathbf{h}(\mathbf{a}))$ will depend on both the update schedule $\mathbf{a}$ and updated forecasts $\mathbf{h}(\mathbf{a})$. In order to make this dependence explicit, rewrite the cost function as the sum of costs over each interval $[a_k, a_{k+1})$. The value of each term in the sum will depend only

on the forecast produced for that interval,

$$h(t) = h(a_k) \text{ for } a_k \leq t < a_{k+1} \tag{7.7}$$

so that for $a_k \leq g_0 < a_{k+1}$, equation 7.6 may be rewritten as:

$$\begin{aligned} C_F(g_0; \mathbf{a}, \mathbf{h}(\mathbf{a})) &= \sum_{i=0}^{k-1} \int_{a_i}^{a_{i+1}} w(t) C(g_0 - h(a_i)) dt \\ &\quad + \int_{a_k}^{g_0} w(t) C(g_0 - h(a_k)) dt \qquad (7.8) \\ &= \sum_{i=0}^{k-1} W(a_i, a_{i+1}) C(g_0 - h(a_i)) \\ &\quad + W(a_k, g_0) C(g_0 - h(a_k)) \qquad (7.9) \end{aligned}$$

where

$$W(a, b) = \int_a^b w(t) dt \tag{7.10}$$

In addition to this forecast cost, a cost K_u is assessed on each update, where K_u is expressed in units that are equivalent to the units used for the forecast error cost. The minimum cost of updating for any flight is $2K_u$: a charge of K_u for the initial forecast at a_0 and another for the update when the aircraft takes off at g_0. However, the update cost for any particular departure time g_0 will depend on the number of update epochs before g_0. The expected number of updates over all g_0 is one result presented in the next subsection.

7.3.2 Expected Cost of a Sequence of Forecasts

The expected forecast error cost over all g_0 is simply the expected value of C_F. The function $f_g(g_0)$ is the prior pdf of random variable g and let $F(\mathbf{a}, \mathbf{h}(\mathbf{a}))$ be the expected forecast error cost under update schedule $\mathbf{a}$ and forecast sequence $\mathbf{h}(\mathbf{a})$. Therefore,

$$F(\mathbf{a}, \mathbf{h}(\mathbf{a})) = \int_{-\infty}^{\infty} C_F(g_0; \mathbf{a}, \mathbf{h}(\mathbf{a})) f_g(g_0) dg_0 \tag{7.11}$$

Using equation 7.9 and rearranging terms, we find:

$$F(\mathbf{a}, \mathbf{h}(\mathbf{a})) = \sum_{k=0}^{n} M_k(h(a_k)) \tag{7.12}$$

where for $k = 0 \ldots n-1$

$$M_k(h(a_k)) = \int_{a_k}^{a_{k+1}} W(a_k, g_0) C(g_0 - h(a_k)) f_g(g_0) dg_0 + W(a_k, a_{k+1}) \int_{a_{k+1}}^{\infty} C(g_0 - h(a_k)) f_g(g_0) dg_0 \quad (7.13)$$

and

$$M_n(h(a_n)) = \int_{a_n}^{\infty} W(a_n, g_0) C(g_0 - h(a_n)) f_g(g_0) dg_0. \quad (7.14)$$

In this sum, each term $M_k(h(a_k))$ represents the cost which, according to the prior distribution, we expect to accumulate during the interval $[a_k, a_{k+1})$. Throughout this interval, $h(a_k)$ is used as the forecast. $M_n(h(a_n))$ is the cost which we expect to accumulate between a_n and ∞ when using the forecast $h(a_n)$. This additivity of costs motivates the dynamic programming approach described in section 7.4.2.

Let F_g be the cumulative distribution function of g. Given update schedule $\mathbf{a}$, the expected number of epoch is:

$$U(\mathbf{a}) = \sum_{k=0}^{n} (1 - F_g(a_k)) + 1 \quad (7.15)$$

The single update added to this total represents the final update when the flight departs. The total expected cost for update schedule $\mathbf{a}$ and the sequence of forecasts $\mathbf{h}(\mathbf{a})$ is $F(\mathbf{a}, \mathbf{h}(\mathbf{a})) + K_u U(\mathbf{a})$. The goal is to solve the minimization problem:

$$\min_{\mathbf{a}, \mathbf{h}(\mathbf{a})} F(\mathbf{a}, \mathbf{h}(\mathbf{a})) + K_u U(\mathbf{a}). \quad (7.16)$$

The minimization problem can be expanded as:

$$\min_{\mathbf{a}, \mathbf{h}(\mathbf{a})} \sum_{k=0}^{n} \{M_k(h(a_k)) + K_u(1 - F_g(a_k))\} + K_u \quad (7.17)$$

7.3.3 Expected Number of Updates Under Discrete Conditional Updating Using Conditional Means

A useful property of the discrete conditional updating procedure is that under certain conditions the expected number of updates does not depend on the mean or variance of the

prior distribution of take-off time. Intuitively, an increase in the variance of the distribution lengthens the time period that must be covered by the schedule, but also widens the gaps between updates, if updates are scheduled according to the conditional means of the distribution.

Let g be distributed according to $F_g(g)$ and let $\mathbf{a} = \{a_0, a_1, \ldots\}$ be the update schedule for g under discrete conditional updating using conditional means. That is:

$$\begin{aligned} a_0 &= -\infty \\ a_1 &= E(g) \\ a_2 &= E(g|g \geq a_1) \\ &\vdots \\ a_i &= E(g|g \geq a_{i-1}) \end{aligned}$$

Here we are assuming that updates may occur anywhere, and are not restricted to five-minute intervals. Also let equation 7.15 define $U(\mathbf{a})$, the expected number of updates under schedule $\mathbf{a}$.

Claim 7.1 *Let $g' = cg + d, c > 0$, and let $\mathbf{a}' = \{a'_0, a'_1, \ldots\}$ be the update schedule for g' under discrete conditional updating. Then:*

$$U(\mathbf{a}') = \sum_{k=0}^{n}(1 - F_g(a'_k)) + 1 = U(\mathbf{a}) \tag{7.18}$$

Proof: We will show by induction that $a'_k = ca_k + d$, and this will imply that $F_g(a'_k) = F_g(a_k)$ for all k. First, $a'_0 = ca_0 + d = -\infty$. Assume that $a'_{k-1} = ca_{k-1} + d$. Then

$$\begin{aligned} a'_k &= E(g'|g' \geq a'_{k-1}) \\ &= E(cg + d|cg + d \geq ca_{k-1} + d) \\ &= ca_k + d \end{aligned}$$

Therefore, for all k, $F_{g'}(a'_k) = p(g' \leq a'_k) = p(cg + d \leq ca_k + d) = F_g(a_k)$.

For example, if g is distributed as a standard normal random variable, then sum 7.15 evaluates to $U(\mathbf{a}) = 2.9$. By the claim above, this is the expected number of updates under

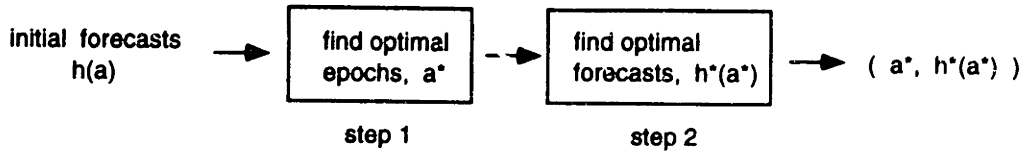


Figure 7-5: Two-step heuristic for finding optimal update epochs and forecasts.

discrete conditional updating for a normally distributed random variable with *any* mean and variance.

7.4 Optimal Update Schedules and Forecasts

An optimal solution to the forecast cost minimization problem requires that expression (7.16) be minimized over both the update schedule $\mathbf{a}$ and the sequence of forecasts $\mathbf{h}(\mathbf{a})$. While an exact solution to this difficult nonlinear optimization problem might be a subject for further study, the decomposition heuristic presented here produces near-optimal results when applied to the numerical examples in the next section. The heuristic divides the problem into two steps, first finding an optimal update schedule under a given method for deriving forecasts and then revising the forecasts to be optimal for the update schedule derived in the first step. The steps are shown in the flow chart, Figure 7-5. Each of these two steps may be used independently. A system may be required to follow a predetermined update schedule, and the second step determines an optimal sequence of forecasts under this constraint. Alternatively, a particular class of forecasts, such as the simple conditional mean, may be preferred at each update epoch, and the first step finds the optimal times at which these forecasts should be produced.

The next subsection discusses the production of forecasts and derives two methods which will be used in the first and second steps of the heuristic. Two classes of forecasts are presented. One is the optimal one-time forecast and the other is the optimal sequential forecast. *One-time forecasts* minimize the expected value of the cost function $C(g - h(t))$ and are optimal in the traditional sense: the cost is evaluated at only one time, i.e. when the aircraft departs. These one-time forecasts will be used to find the optimal update epochs in the first step of the heuristic. *Optimal sequential forecasts* minimize the expected cumulative cost defined in equation (7.11) for a given update schedule; step 2 in the heuristic solves for the optimal sequential forecasts.

Subsection 7.4.2 describes a dynamic program for determining the optimal update schedule, given a particular method for making forecasts. This is the algorithm used in step 1 of the heuristic. The dynamic program assumes that we have a finite number of potential update epochs, $t_0, t_1, \ldots, t_m$. For the ETMS, the t_k might be the beginning of each five minute cycle in a large window around the proposed departure time. The dynamic program finds the update schedule which minimizes the expected total cost, given forecasts for each potential update epoch. Finally, subsection 7.4.4 describes the two-step heuristic in more detail.

7.4.1 Optimal One-Time and Sequential Forecasts

Assume that at potential update epoch t_k we may make a single updated forecast and that we must stick with this forecast until the flight departs. The optimal one-time forecast, $h^o(t_k)$, minimizes:

$$E\left[C(g - h(t_k))|g > t_k)\right] = \int_{t_k}^{\infty} C(g_0 - h(t_k)) \frac{f_g(g_0)}{1 - F_g(t_k)} dg_0 \tag{7.19}$$

A simple expression for $h^o(t)$ can be derived if the cost function is piecewise linear. The optimal one-time forecast is:

$$h^o(t_k) = F_g^{-1}\left[\frac{K_l(1 - F_g(t_k))}{K_l + K_r} + F_g(t_k)\right]. \tag{7.20}$$

Note that if $t = -\infty$ and $K_l = K_r$, then $h^o(t)$ is the median of the distribution.

For a particular update schedule $\mathbf{a}$, the one-time forecasts will not always minimize $F(\mathbf{a}, \mathbf{h}(\mathbf{a}))$. Let the sequence $\mathbf{h}^*(\mathbf{a})$ be the optimal sequential forecasts, so that $\mathbf{h}^*(\mathbf{a})$ is the sequence of optimal forecasts $\{h^*(a_0), h^*(a_1), \ldots, h^*(a_n)\}$ which minimizes $F(\mathbf{a}, \mathbf{h}(\mathbf{a}))$. From equation 7.12, the kth term in the sum of forecast error costs depends only on $h(a_k)$. Therefore, each element of $\mathbf{h}^*(\mathbf{a})$, $h(a_k)$, must minimize the term $M_k(h(a_k))$. Appendix 7A presents a simple procedure for finding the optimal sequential forecast $h^*(a_k)$.

7.4.2 Optimal Update Schedules Given Forecasts

Assume that we are given forecast $h(t_k)$ at each potential update epoch $t_k, k = 0, 1, \ldots, m$. The forecast $h(t_k)$ may be the conditional mean $E(g|g > t)$ or it may be the more general

one-time forecast described in the previous subsection. Assume that no forecast updates are possible after t_m (the time of this last potential update epoch can be chosen so that $p(g > t_m)$ is negligible and any updating after t_m would not be expected to reduce forecast costs). In order to simplify the notation, we define $t_0 \equiv -\infty$, so that $h(t_0)$ is our initial, prior forecast. The following dynamic program finds the optimal update schedule:

Indices: $k = 0, 1, \ldots m$, the index for each potential update epoch.

Potential update epochs: t_k.

State variables:

$$x_k = \text{index of the most recent update epoch}$$
$$(x_k = 0, 1, \ldots k-1)$$

$$d_k = \begin{cases} 0 & \text{if the aircraft has not departed by time } t_k \\ 1 & \text{if the aircraft has departed by time } t_k \end{cases}$$

Control Variables: Define a zero-one variable u_k to represent the decision to update at epoch t_k.

$$u_k = \begin{cases} 0 & \text{if no update occurs at } t_k \\ 1 & \text{if an update occurs at } t_k \end{cases} \tag{7.21}$$

for $1 \leq k \leq m$ and let $u_0 = 0$.

Transition Probabilities: The state variable d_k is stochastic with transition probabilities:

$$\begin{aligned} p(d_k = 0 | d_{k-1} = 0) &= \frac{1 - F_g(t_k)}{1 - F_g(t_{k-1})} \\ p(d_k = 1 | d_{k-1} = 0) &= \frac{F_g(t_k) - F_g(t_{k-1})}{1 - F_g(t_{k-1})} \\ p(d_k = 0 | d_{k-1} = 1) &= 0 \\ p(d_k = 1 | d_{k-1} = 1) &= 1 \end{aligned}$$

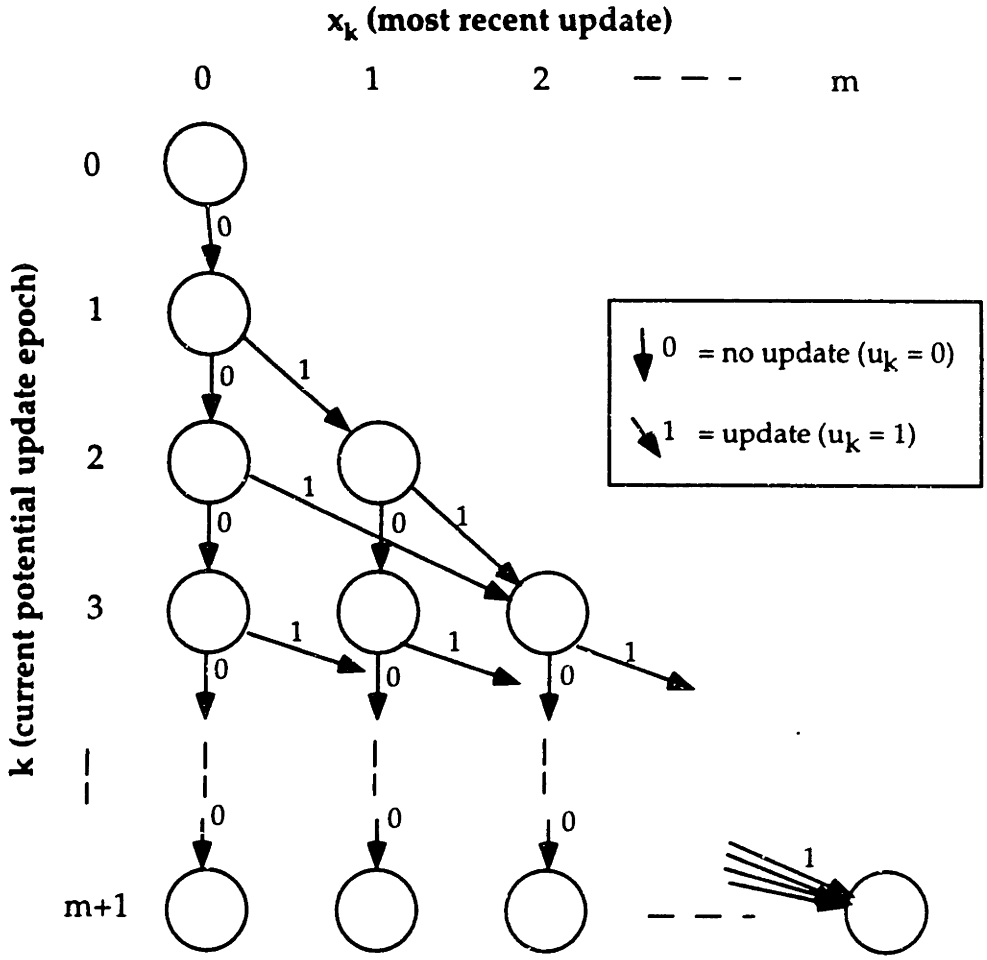


Figure 7-6: State Space of x_k

The state variable x_k is deterministic. Its value is determined by the value of the control variable u_k:

$$x_{k+1} = \begin{cases} x_k & \text{when } u_k = 0 \\ k & \text{when } u_k = 1 \end{cases} \tag{7.22}$$

The algorithm is initialized by setting $x_0 = 0$. Figure 7-6 illustrates the state space of x_k. The variable allows the dynamic program to keep track of the time since the last update.

Costs: In order to utilize dynamic programming, the expected cost after time t_k must be expressed in terms of the sum of the expected cost between t_k and t_{k+1} and the expected cost after t_{k+1}. With such a cost function, the optimal control law may be found by working backwards from t_m and choosing the values of u_k that minimize the sum at each stage.

Let $J_k(x_k), k = 0, \ldots, m$ be the expected remaining cost after time t_k, given that the last

update occurred at time t_{x_k}. Remember that x_k may take on values from 0 to $k-1$. At time t_k an update may occur and the value of x_{k+1} is determined according to equation (7.22). The forecast $h(t_{x_{k+1}})$ applies during the period $[t_k, t_{k+1})$, since the update occurs at the beginning of the period. If $d_k = 1$, the expected remaining cost is zero since the aircraft has departed and no more forecasts are necessary. If $d_k = 0$, the expected remaining cost is the sum of the expected forecast error cost from t_k to t_{k+1} with forecast $h(t_{x_{k+1}})$, the update cost at t_k if there is an update, and the expected costs accumulated after t_{k+1}. The three terms in the following expression represent these three expected costs:

$$J_k(x_k) = M_k(h(t_{x_{k+1}})) + u_k K_u(1 - F_g(t_k)) + J_{k+1}(x_{k+1}) \tag{7.23}$$

Define the terminal cost $J_{m+1}(x_{m+1}) = K_u$ and work backwards to $k = 0$ to find that the total expected cost is the following sum:

$$J_0(x_0) = \sum_{k=0}^{m} \{M_k(h(t_{x_{k+1}})) + u_k K_u(1 - F_g(t_k))\} + K_u \tag{7.24}$$

Note that this is similar to the expected total cost in equation (7.17), with the $h(t_k)$ replaced by $h(t_{x_{k+1}})$, the control variable u_k multiplying the potential update costs, and the sum moving over all m potential update epochs.

Since the costs in equation (7.24) are additive, the optimal control law may be determined by a dynamic program. Figure 7-6 demonstrates that when there are m potential update epochs, the dynamic program must calculate and compare costs along a pair of arcs from $\frac{m(m+1)}{2}$ nodes. Therefore, the algorithms has a running time of $O(m^2)$. Each step requires the calculation of $M_k(h(t_{k+1}))$, and the computational requirements of this calculation depend on the form of $w(t)$, $f_g(g_0)$, and the cost function for forecast errors.

If $w(t)$ is constant within each time period and the costs of forecast errors are linear, a simple expression for M_k can be found. This expression is derived in Appendix B of this chapter. Appendix B also derives algebraic expressions for M_k for the cases when $f_g(g_0)$ is a normal or gamma distribution. These algebraic expressions will enable efficient computations of M_k for the empirical experiments in Section 7.5. Evaluating M_k may require numerical integration for other forms of $w(t)$, $f_g(g_0)$, or the cost function.

7.4.3 Dynamic Programming Solution with Required Updates

While the dynamic programming solution minimizes the expected cost of the update schedule, it may produce solutions that seem foolish to the users. If the cost of an update is high, this method may stick with a forecast even after the time of that forecast has come and gone. For example, at 30 min. the method may produce a forecast of 37 min. and may schedule the next update epoch for 50 min. Therefore, update epochs at 40 and 45 min. pass without an update, even though the forecast of 37 min. is obviously incorrect.

While such an update schedule may be globally optimal, it will not encourage faith in the system. One may modify the dynamic programming method to require an update if the last forecast has expired, i.e. if $a_k > h(a_k)$. This modified dynamic program with 'required updates' was implemented along with the original dynamic program. However, in the following computational experiments, the solution produced by this algorithm was essentially identical to the solution produced by the original dynamic programming algorithm. In general, the original dynamic program did not allow the forecasts to expire. Therefore, results from the 'required updates' algorithm will not be displayed.

7.4.4 A Heuristic for Combining Schedules and Forecasts

The procedure in section 7.4.1 finds optimal forecasts, *given an update schedule*, while the dynamic program generates update schedules. The following heuristic combines update schedules with optimal forecasts. As shown in Figure 7-5, the heuristic first builds update schedules with the forecasts $h^o(t_k)$ and then finds the optimal forecasts, given the schedule:

Step 1: Use the dynamic program to find an update schedule, given the one-time forecasts $h^o(t_k)$ at each potential update epoch.

Step 2: Find optimal sequential forecasts for the update schedule generated in Step 1.

The heuristic uses the one-time forecasts in Step 1, for these forecasts are computationally easy to generate. Comparing costs incurred by forecast sequences from Steps 1 and 2 will demonstrate the increase in forecast accuracy after re-solving for the optimal forecasts.

However, the dynamic program is not the only possible method for generating update schedules in Step 1. An alternative method, discrete conditional updating, was described in section 7.2. Whether a schedule is generated with the dynamic program or by discrete

conditional updating, forecast accuracy can be improved by resolving for the optimal forecasts. In the numerical example of the next section we will use the discrete conditional updating scheme with one-time forecasts, in Step 1, to generate an update schedule. As in step 2 of the heuristic, the resulting update schedule will be used to solve again for optimal sequential forecasts. Comparing the results from the dynamic program with those produced by discrete conditional updating will suggest the benefits of having an optimal schedule.

7.5 Examples and Numerical Results

In this section numerical examples illustrate some of the differences among the procedures that have been presented. Six distinct methods are used to generate forecast sequences: constant interval updating, discrete conditional updating with one-time forecasts and with optimal sequential forecasts, dynamic programming with one-time forecasts and with optimal sequential forecasts, and continuous conditional updating. Each procedure has received a general description in previous sections. This section describes the parameters of the numerical examples as well as a few of the implementation details of the updating algorithms. Then the update schedules, forecasts, and expected costs produced by each procedure are compared. The sensitivity of the procedures to two parameters is also tested. The dynamic programming technique is tested over a range of update costs (K_u), and all procedures are compared over a range in the variance of take-off times.

The first set of experiments described below are conducted with forecasts that are normally distributed. A second set, presented at the end of the section, uses a prior gamma distribution. Recall from Section 7.1 that forecasts produced by the roll-out time models roughly follow a gamma distribution.

7.5.1 Parameters for the Numerical Examples

The first numerical example models an air traffic management situation in which departure delays are large, with a mean of sixty minutes and a standard deviation of forty minutes. Such large and variable delays are usually a product of the unsettled weather that may disrupt some airports for a few hours each month. Accurate departure time forecasts are particularly important during these times, since the capacities of the runways and the nearby sectors (air traffic control regions) are often reduced by poor weather.

In this first example the take-off time distribution of a particular flight on a bad-weather day is normal with a mean of one hour after the scheduled departure time and a standard deviation of 40 minutes, so that $g \sim N(60, 40^2)$. With this probability distribution it is possible for an aircraft to depart before its scheduled departure time. This was seen in the data, particularly among flights by general aviation and freight carriers.

One half hour before the expected take-off time of this aircraft, (equivalently, one half hour after the original scheduled departure time), regional air traffic managers begin to include the flight in their sector congestion forecasts. The departure time forecast will be used in the congestion calculations until the flight takes off (after take-off, a forecast is no longer necessary). Therefore, forecast errors begin to be consequential one-half hour after the scheduled departure time and continue to have an effect until take-off. This suggests a forecast error weighting function which rises from zero to one at 30 min. and remains at one thereafter:

$$w(t) = \begin{cases} 0 & \text{if } t < 30 \\ 1 & \text{if } t \geq 30 \end{cases} \tag{7.25}$$

$$\equiv \delta(t - 30) \tag{7.26}$$

In addition, suppose that the air traffic manager is indifferent between late and early forecasts and has a piecewise linear cost function for forecast errors, so that $C(g_0 - h(t)) = |g_0 - h(t)|$. Finally, $K_u = 25$, so a penalty of 25 units is assessed for each update. Subsection 7.5.3 will discuss how an appropriate update penalty may be determined.

7.5.2 Comparison of Forecast Sequences

Six procedures produced update schedules and forecasts from these parameters. The first five procedures are restricted to potential update epochs at five-minute intervals, i.e. at 0 min., 5 min., 10 min. This conforms to the actual cycle time of the ETMS. The sixth technique, continuous conditional updating, assumes that the cycle time is zero and that updates occur at all times. The six procedures are:

1. *Constant interval updating* with 5-minute intervals. This procedure is described in section 7.2 above.

2. *Discrete conditional updating using one-time forecasts.* This corresponds to Step 1 in the heuristic with discrete conditional updating to set the timing of the update epochs and with a one-time forecast, $h^o(t)$, produced at each update epoch.

3. *Dynamic programming using one-time forecasts.* This is also Step 1 in the heuristic, but with the dynamic program (DP) generating an update schedule.

4. *Discrete conditional updating using optimal sequential forecasts.* Discrete conditional updating determines the update schedule and then optimal forecasts are generated for that schedule. These procedures correspond to steps 1 and 2 in the heuristic of Figure 7-5.

5. *Dynamic programming using optimal sequential forecasts.* The DP determines the update schedule and then optimal forecasts are found, given this schedule. Again, this procedure corresponds to steps 1 and 2 in the heuristic, using the DP in step 1 rather than discrete conditional updating.

6. *Continuous conditional updating.* We adapt the procedure described in section 7.2. Instead of producing continuous updates with conditional means, we use the conditional median, which minimizes the piecewise linear cost function.

Figure 7-7 displays the update schedules and forecasts produced by the first three procedures described above. The vertical axis of the figure is the predicted time remaining for each sequence of forecasts, so that if $h(t)$ is the forecast at time t, then the graph shows $h(t) - t$ plotted against t. Update epochs are visible as vertical jumps in the predicted remaining time. The figure displays the update *schedules*, and for any particular flight, the sequence of forecasts will be cut short by the actual departure. The figure only displays the first ninety minutes of the schedule, but the forecast sequence was calculated to two hours after the scheduled departure time. Under the probability distribution of g, flights are extremely unlikely to depart after two hours.

Constant interval updating begins with a forecast equal to the prior mean of 60 min., so that the time remaining is 35 min. at $t = 25$. An update occurs at 60 min. and at 5 minute intervals thereafter. This produces the dashed, jagged line at the bottom of the graph. Discrete conditional updating begins with an initial forecast of 60 min. and first revises this forecast at 60 min. The revised forecast is 87 min. This is the one-time forecast, a

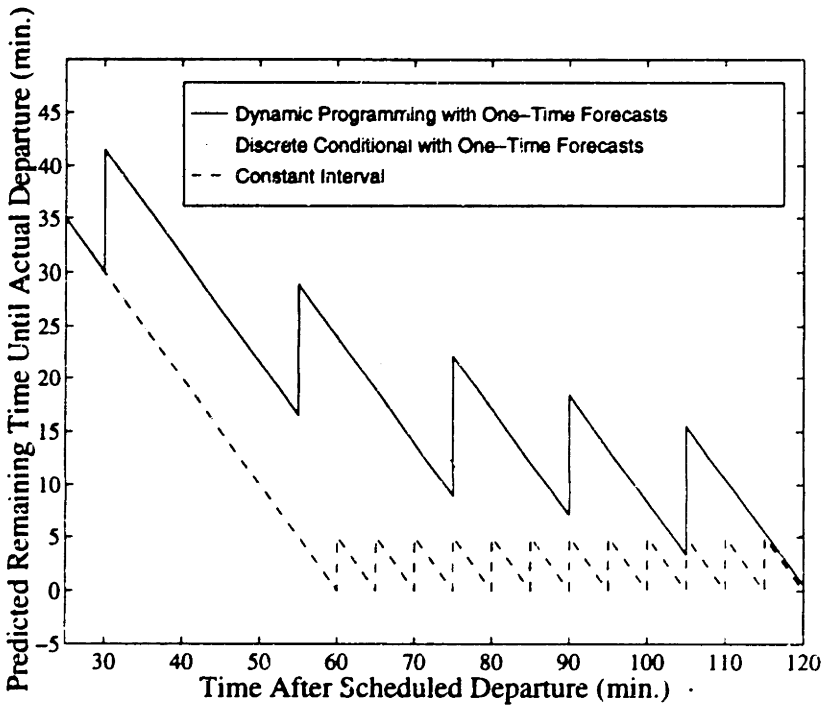


Figure 7-7: Predicted time remaining for three forecast sequences with parameters $g \sim N(60, 40^2)$, $w(t) = \delta(t-30)$, $K_r = K_l = 1$, and $K_u = 25$.

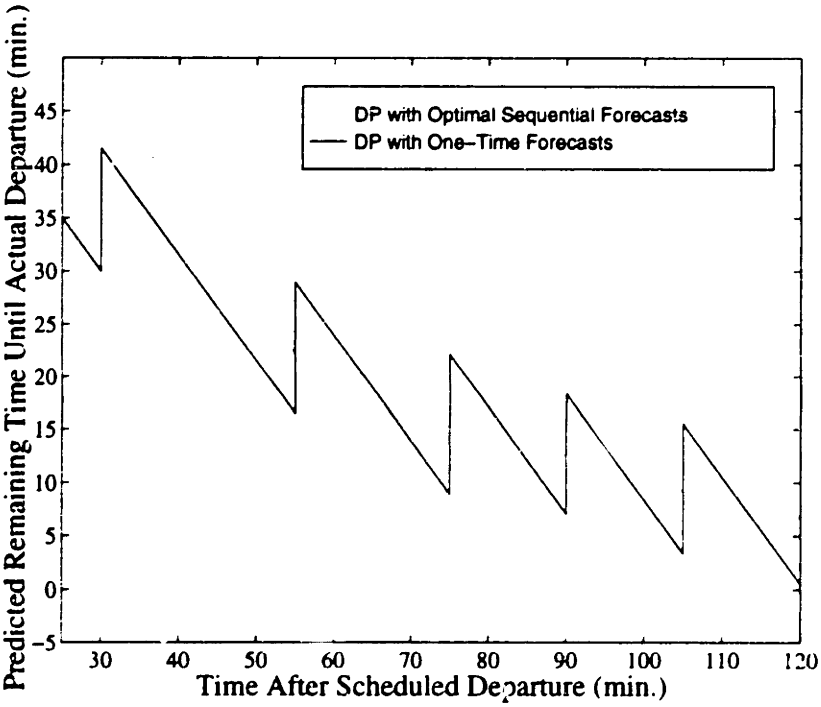


Figure 7-8: Predicted time remaining for forecast sequences generated by dynamic program with optimal sequential and one-time forecasts. Parameters are $g \sim N(60, 40^2)$, $w(t) = \delta(t-30)$, $K_r = K_l = 1$, and $K_u = 25$.

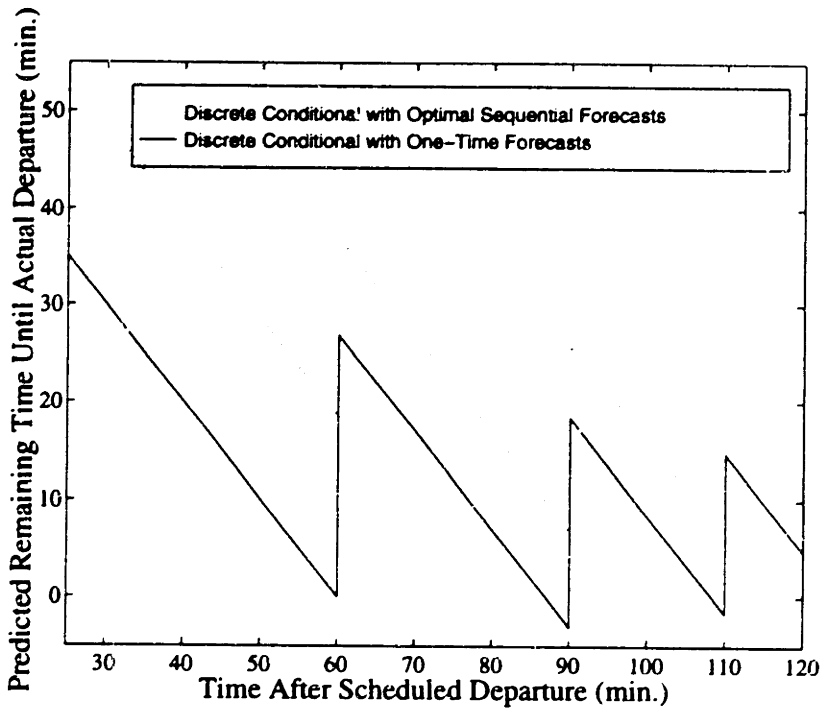


Figure 7-9: Predicted time remaining for forecast sequences generated by discrete conditional updating with optimal sequential and one-time forecasts. Parameters are $g \sim N(60, 40^2)$, $w(t) = \delta(t - 30)$, $K_r = K_l = 1$, and $K_u = 25$.

conditional median conditioned on the event $g > 60$. Therefore, at $t = 60$ the predicted time remaining is 27 min. The next update occurs at 90 min., the beginning of the first cycle after the forecast has expired.

The dynamic program also has an initial forecast of 60 min., but prescribes an initial update at 30 min. Recall that the weighting function $w(t)$ is first nonzero at 30 min. The DP strategically places updates so that the sum of expected update and forecast costs are minimized. An update before 30 min. does not help reduce forecast costs, since the weighting function $w(t)$ is zero for $t < 30$ min. On the other hand, if the first update occurs after 30 min. then the obsolete forecast of 60 min. would remain and forecast errors would accumulate. Therefore, 30 minutes is an opportune time for an update. The DP also places subsequent updates at 55, 75, 90 and 105 minutes, times at which forecasts are sufficiently obsolete to justify the cost of an update.

For the discrete conditional procedure, updates are scheduled for 60, 90, and 110 minutes. Under this procedure, most flights will see two, three, or four updates: an initial forecast, updates at 60 and 90 if the flight has not taken off, and a final update when takeoff occurs. The expected number of updates for this procedure is 2.9. While the discrete

conditional procedure prescribes fewer updates than the other two algorithms, we shall see that it produces higher expected forecast costs than the DP solution.

The forecasts produced by the first three procedures are one-time forecasts, which are not optimal for the given update schedules. Figures 7-8 and 7-9 compare the optimal sequential forecasts produced for the discrete conditional and DP procedures with the one-time forecasts produced for these same procedures. The optimal sequential forecasts are consistently higher than the one-time forecasts, about 5 min. higher under the DP update schedule and 10 - 20 min. higher under the discrete conditional updating schedule. This is a consequence of the cumulative nature of the forecast cost function, expression (7.6). The optimal sequential forecast $h^*(a_k)$ produced at update epoch k must minimize the expected cumulative forecast cost for departures after time a_k. Flights which depart immediately after a_k will have little time to accumulate error before take-off, and therefore receive less weight in the function to be minimized. Flights which depart after a_{k+1}, the next update epoch, will accumulate forecast error over the entire period $[a_k, a_{k+1})$, and therefore will have a relatively large impact on the cost function. This shifts the forecast $h^*(a_k)$ towards later flights. However the difference between one-time and optimal sequential forecasts decreases as the distance between update epochs decreases. For example, the difference between forecasts is smaller for the DP update schedule than for the discrete conditional updating schedule, since the DP schedule has more frequent update epochs. As the distance between update epochs approaches zero, as in continuous conditional updating, the one-time and optimal sequential forecasts converge.

The expected forecast error costs, $F(\mathbf{a}, \mathbf{h}(\mathbf{a}))$ and update costs $K_u U(\mathbf{a})$ under each forecast sequence are displayed in figure 7-10. The number in parentheses at the top of each bar is the expected number of updates, $U(\mathbf{a})$. The horizontal bar at 665 minutes indicates the expected forecast error cost if continuous conditional updating were used. Given the distribution assumed for take-off time, this is a lower bound on the expected forecast error cost. Constant interval updating has the largest expected forecast and update costs, due to its frequent updates and uninformative forecasts. As was seen in figure 7-7, the discrete conditional procedure is relatively "stingy" with its updates, having only 2.9 expected updates to the dynamic program's 4.2. However, when using the one-time forecasts, the dynamic program achieves a significantly lower expected forecast error cost than discrete conditional updating with one-time forecasts.

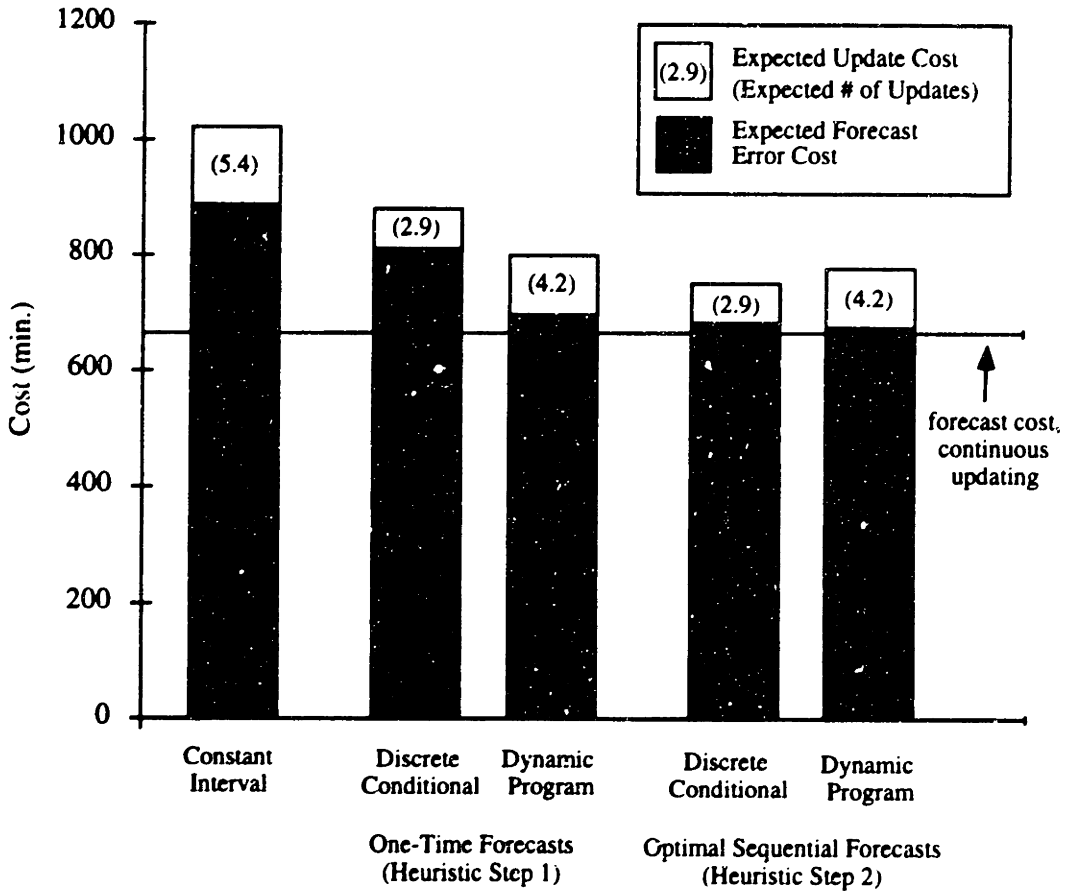


Figure 7-10: Expected forecast error costs (dark bars) and update costs (white bars) for six update schedules. The expected number of updates under each schedule is shown in parentheses.

When optimal sequential forecasts are used with each update schedule, the procedures come close to the lower bound on expected forecast error cost. The dynamic programming schedule, when combined with optimal forecasts, has an expected forecast error cost that is only 1.2% higher than the lower bound, while the discrete conditional schedule produces expected forecast error costs that are 2.5% higher than the lower bound. The latter schedule, with its small number of expected updates, achieves the lowest overall cost. Its expected total cost is 26% lower than the 5-minutes constant interval procedure, which is currently implemented by the ETMS.

7.5.3 Sensitivity to Update Cost and Prior Standard Deviation

In this section we examine the relationship between forecast error cost and the number of updates. As the update penalty increases, the number of updates scheduled by the dynamic program decreases while the expected forecast error cost increases. Figure 7-11 displays this curve for one-time forecasts (heuristic step 1) and optimal sequential forecasts (heuristic step 2). The lower bound on the expected number of updates is two, since both the initial forecast and a final update are required. The points generated by $K_u = 25$ and other update costs are marked on the graph. The curve for the DP-generated update schedule with optimal sequential forecasts is not monotonically decreasing, since the combination of update schedules and forecasts may not be globally optimal for a particular K_u.

An analyst may use this curve to determine the update cost which produces a desired level of forecast accuracy for a reasonable number of expected updates. The graph demonstrates that for this distribution of g and when optimal sequential forecasts are used, the forecast error cost is relatively insensitive to the number of updates. In general, the sensitivity will depend on both the distribution of g and the weighting function $w(t)$.

Figure 7-12 displays the forecast costs under each procedure over a range in the prior standard deviation of g. Constant interval and discrete conditional (DC) updating are most sensitive to increases in the standard deviation. However, with optimal sequential forecasts, both the dynamic program and discrete conditional updating remain close to the lower bound set by the continuous conditional updating.

Figure 7-13 displays the expected number of updates under the three scheduling procedures as the prior standard deviation increases. The expected number of updates under the constant interval procedure rises linearly with the prior standard deviation, while the

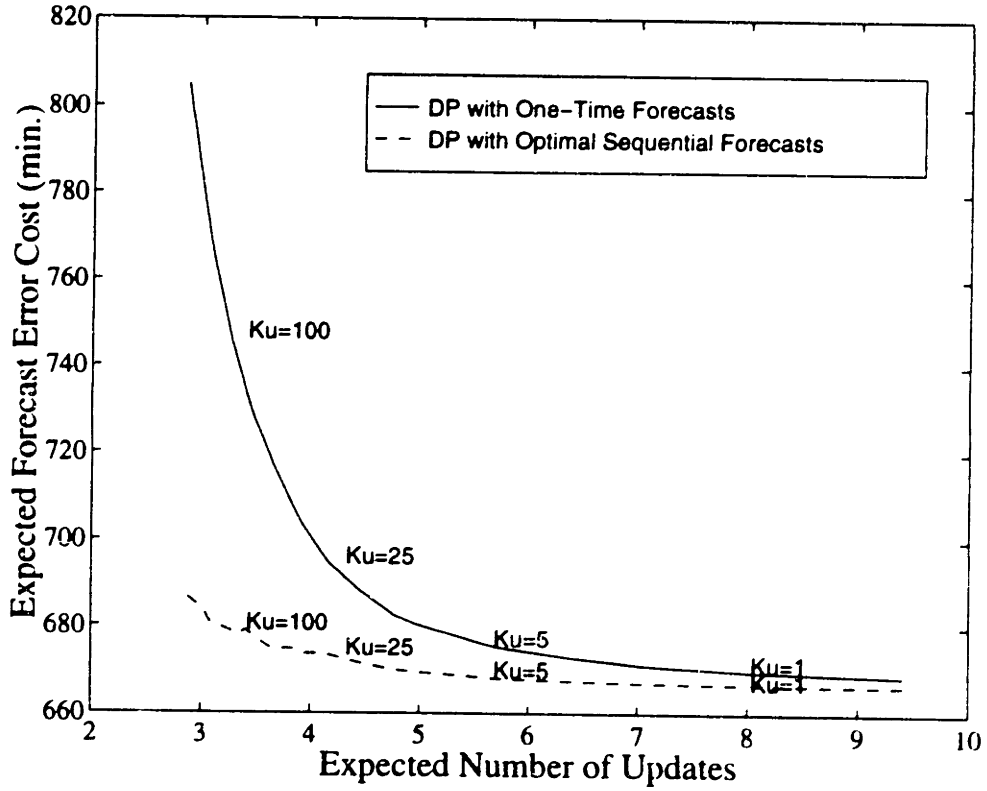


Figure 7-11: Expected Number of Updates and Forecast Costs Produced by the DP Over a Range of Update Costs.

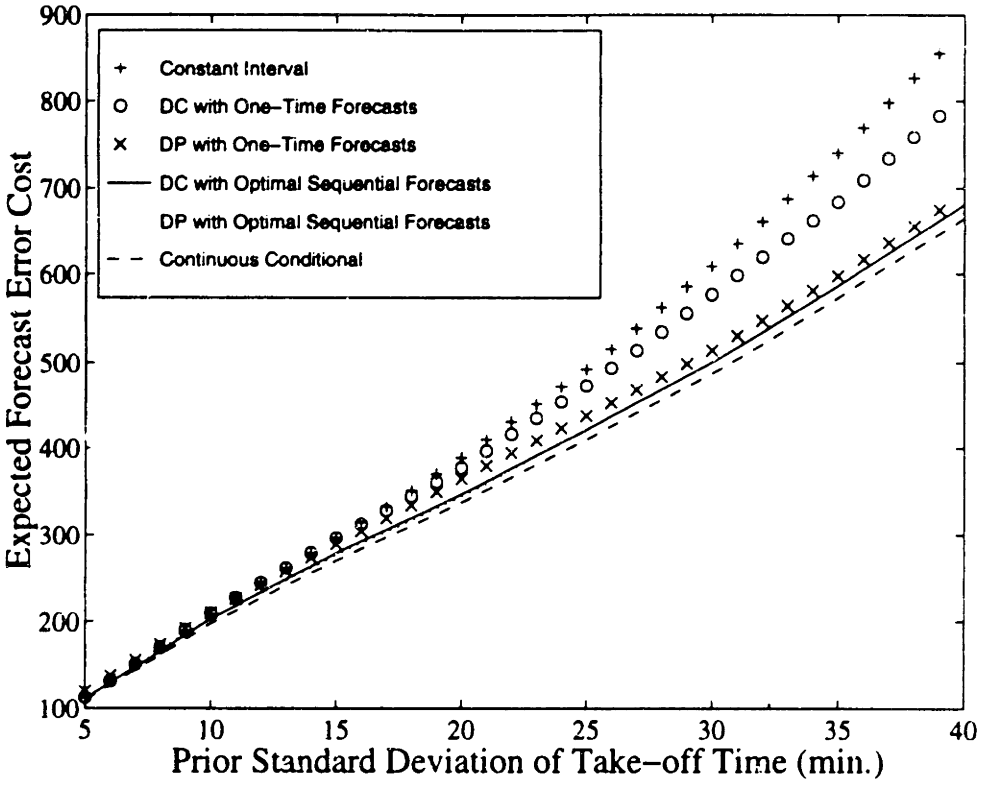


Figure 7-12: Expected forecast error cost with increasing standard deviation of the prior distribution (σ). Parameters are $g \sim N(60, \sigma)$, $w(t) = \mu_1(30)$, $K_r = K_l = 1$, and $K_u = 25$.

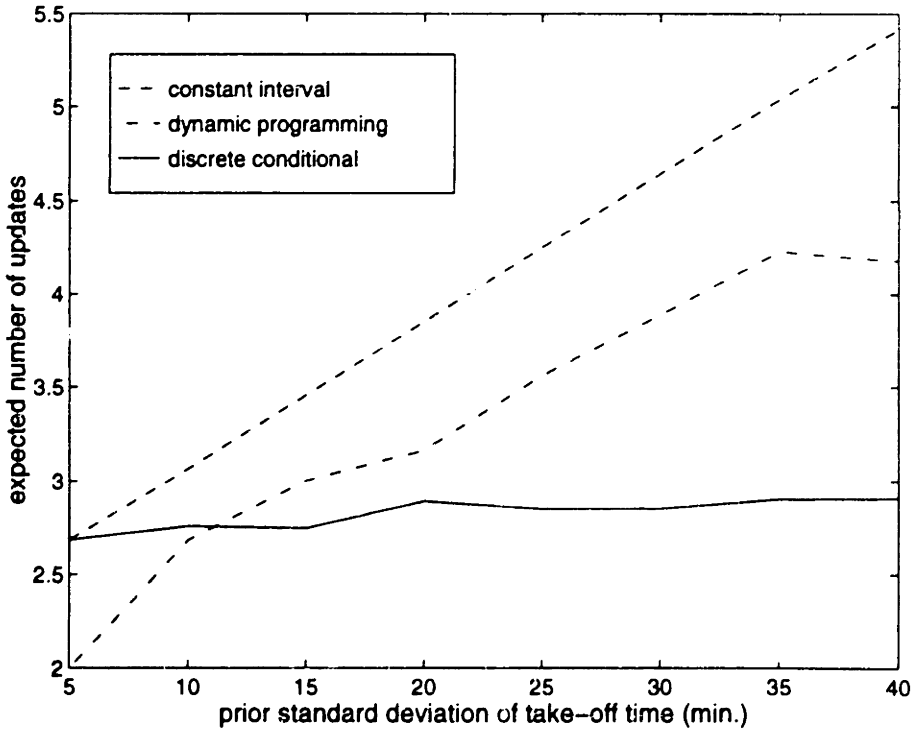


Figure 7-13: Expected number of updates with increasing standard deviation of the prior distribution (σ). Parameters are $g \sim N(60, \sigma)$, $w(t) = \mu_1(30)$, $K_r = K_l = 1$, and $K_u = 25$.

number under the discrete conditional procedure is almost constant. The dynamic program finds a middle ground between these two, balancing update costs with forecast accuracy.

7.5.4 Performance of the Algorithms with ETMS Data

In Chapter 3, we saw that the standard deviation of ground transit times at Logan Airport is between fifteen and twenty minutes (see Table 3.2). Figure 7-12 shows that in this range, there is little difference between the performances of the update schedules in terms of forecast accuracy, while Figure 7-13 shows that the expected number of forecast updates under each procedure also fall within a narrow range. These experiments assume that GTTs are distributed as a normal random variable. However, in Section 7.1 we found that the distribution of forecasts is more closely approximated with a Gamma distribution. Computational experiments again find that there is little difference between procedures when the forecasts follow a Gamma distribution with the parameters estimated in Section 7.1. Expected total costs (forecast error plus update costs) under the dynamic program's update schedule were only 3% less than the forecast costs under the constant interval procedure. The constant interval procedure performed particularly well in this case, for the optimal conditional forecasts at each five-minute interval were close to the constant, five-minute forecasts produced by the constant interval procedure. In general, the choice of an update procedure is not an important consideration when the prior forecast is reasonably accurate. The utility of the optimal updating procedures is most apparent when forecast variance is high. However, it is precisely at these times that forecasts may matter most.

In addition, these experiments may have underestimated the value of the optimal forecasting procedures, for we have assumed that the constant interval procedure begins with the same prior mean used by both the discrete conditional procedure and dynamic program. Suppose instead that the ETMS predicts a ground transit time of 20 minutes for all flights, so that the five-minute updates begin at 20 min. and continue thereafter until the flight departs. This update schedule is shown in the second plot of Figure 7-14. Now suppose that an actual flight has a mean ground transit time of 30 min. and that our forecast has the Gamma error distribution derived in Section 7.1 ($a = 6.5, b = 2.4$). This distribution of GTT is shown in the top plot of Figure 7-14. The third and fourth plots show the discrete conditional and dynamic programming solutions to the update problem (the DP was again solved with an update penalty $K_u = 25$). The first update for the discrete conditional pro-

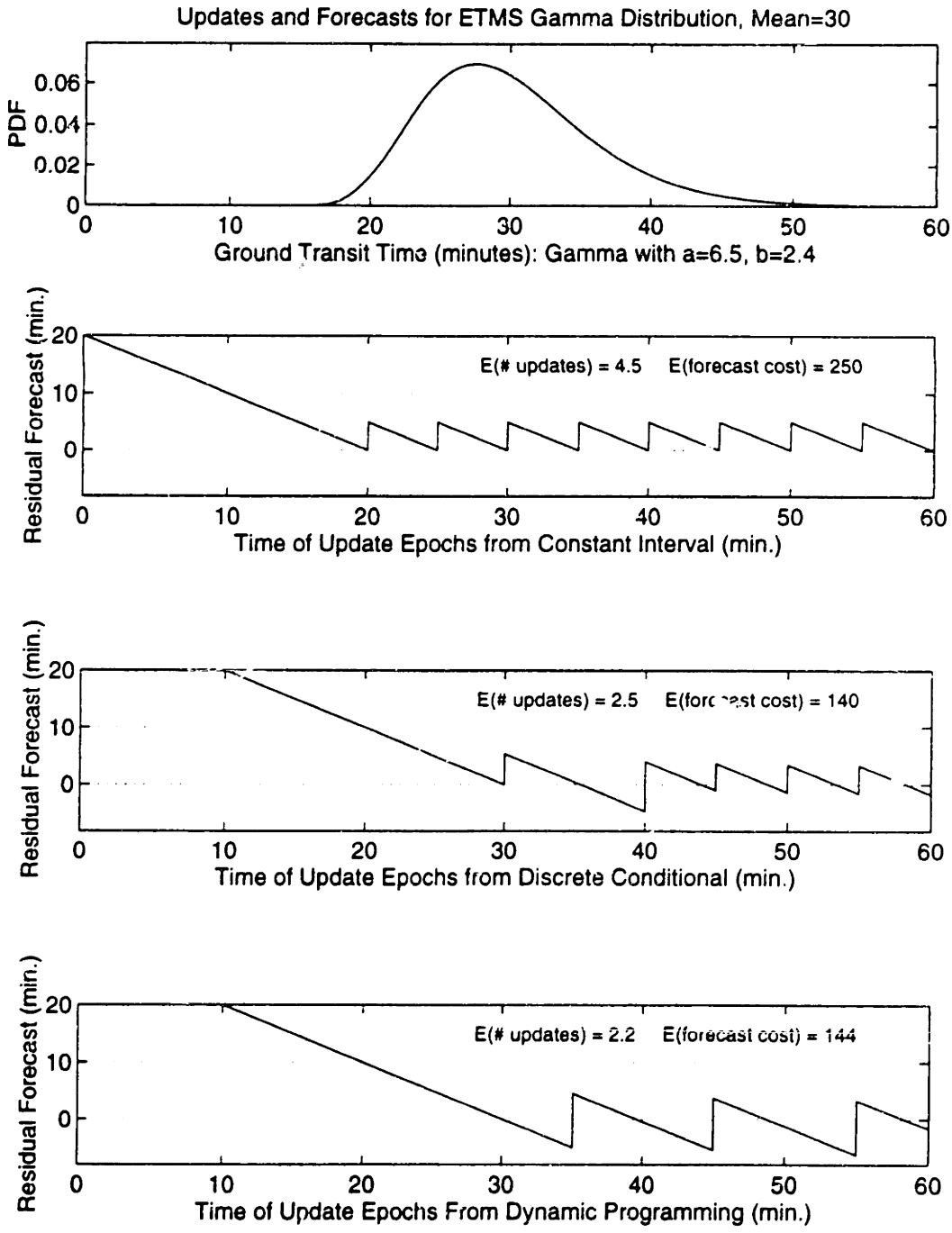


Figure 7-14: Schedule for updates given Gamma distribution fitted to ETMS data.

cedure is at the prior mean, 30 min., and subsequent updates occur at five-minute intervals beginning at 40 minutes. The DP solution schedules updates every ten minutes, beginning at 35 minutes.

The expected forecast and update costs under the discrete conditional and DP procedures are very close. The DP, on average, will see fewer updates than the discrete conditional schedule (2.2 and 2.5 updates, respectively) but will incur a higher forecast cost on average (144 and 140 minutes). Both procedures are much more efficient than the constant interval procedure, which expects 4.5 updates and a forecast cost of 250. In this example, most of the disadvantage of the constant interval procedure is due to the inaccurate initial forecast of 20 minutes. If the prior mean of 30 minutes were used by the constant interval procedure, its update schedule would be similar to that of the discrete conditional procedure (with one more update at 35 minutes), and its forecast and update costs would be within a few percentage points of the other procedures. The ultimate success of any update procedure depends on the accuracy of the prior forecast.

7.6 Conclusions and Extensions

The traditional forecasting problem has been reformulated to allow for the generation of a sequence of forecasts rather than a single forecast. Through the weighting function $w(t)$, the cost function allows an analyst to specify when forecast accuracy is most important. We presented a method for finding optimal sequential forecasts, given a fixed update schedule. A dynamic program found the optimal update schedule, given a particular forecast technique. In the numerical examples, the use of optimal sequential forecasts with both the dynamic program and with the simpler discrete conditional updating procedure obtained forecast errors that were close to the lower bound established by continuous conditional updating.

In general, the optimal sequential forecasts are larger than the one-time forecasts. In the numerical example, the first optimal sequential forecast under the discrete conditional update scheme is almost 20 minutes larger than the prior mean. This bias towards later flights may be removed by heavily penalizing forecasts that result in early, 'surprise' arrivals. This is accomplished by making C_r larger than C_l in the cost equation (7.5) (in the numerical example, the two were equal). Alternatives to the cumulative cost function proposed in equation (7.6), such as an average forecast error rather than an accumulated forecast error,

may also be explored.

By extending the results on optimal inspection schedules presented by Parmigiani [28], further research might find a globally optimal combination of update schedules and forecasts. However, the utility of such an effort is doubtful, given the near-optimal performance of the heuristics presented here. Further research may also find an optimal schedule from a class of simple update schedules. For example, it may be necessary to prescribe an update every L minutes, and procedures for finding the optimal value of L may be derived.

The forecast updates presented in this chapter were driven by the simplest information, the fact that the event of interest had not yet occurred. This allowed us to 'slice' portions of the prior probability density without changing the underlying shape of the distribution. However, it is more realistic to incorporate information from a variety of data sources, and this information may change the shape of the conditional distribution. For example, word that an aircraft is near the head of the runway and 'second to take-off' should reduce the probability that the aircraft's GTT will be in the far right tail of the distribution. A more general dynamic programming approach would take multiple scenarios into account and incorporate many sources of data into the forecasts.

One source of data which may be available soon, and could easily be incorporated into these update procedures, is real-time gate push-back information. The ETMS would construct schedules and forecasts for both gate departure delays and roll-out times and would update the gate departure delay predictions until push-back occurs. Then roll-out predictions would be updated until take-off. Since gate departure delay and roll-out times are approximately independent, a constant roll-out prediction may be used during the gate departure delay. Only after push-back does the roll-out forecast change.

Another important aspect of the departure process which could be incorporated into these algorithms is the possibility of a flight cancellation. In the ETMS, if no cancellation is received, forecast updates are made until an hour passes. This accounts for a significant number of unnecessary updates. It would not be difficult to include cancellation predictions into the optimal update algorithms. One would have to specify a prior probability of cancellation. Also necessary are the costs of false positives and false negatives: the cost of labeling a flight 'canceled' if it does eventually take off, and the cost of predicting a take-off for a flight that is actually canceled. Given these parameters, as well as those already described in this chapter, one may calculate the probability that a flight has been canceled

at any time t, as well as the expected costs of take-off and cancellation predictions.

In general, the empirical results demonstrate increasing advantages for the optimal updating procedures as the level of take-off time uncertainty grows, but the average level of uncertainty in take-off times may not justify the use of the more complex procedures under normal circumstances. The simple discrete conditional procedure achieves a forecast accuracy which is close to the optimal cost, and often uses fewer updates. In Section 7.3.3 we showed that the number of updates performed under the normal distribution by the discrete conditional procedure does not grow with the size of the prior variance, and this is also shown for the numerical example in Figure 7-13.

The tenet behind the discrete conditional update procedure is that forecasts should be based on the *conditional* probability distribution, calculated after incorporating all available information. The largest gains in forecast accuracy may be achieved by collecting and incorporating the most relevant information. This should include real-time status reports about mechanical difficulties, baggage problems, and passenger or crew delays. Currently the only information available is the *absence* of an observed take-off, and the procedures in this chapter do the best they can with this limited information. More information, properly handled, would greatly improve real-time forecasts.

Appendix 7A: Optimal Sequential Forecasts for a Given Update Schedule

We wish to minimize $M_k(h(a_k))$ over $h(a_k)$. To simplify the notation slightly, consider minimizing $M_k(h)$ over forecasts h, and let the optimal solution be designated h^*. Assume that $h \geq a_k$, since h is the forecast of g, conditioned on $g \geq a_k$. The functional form of $M_k(h)$ changes at $h = a_{k+1}$, so there will be two cases to evaluate:

case (i) $(a_k \leq h < a_{k+1})$:

$$\begin{aligned} M_k(h) &= \int_{a_k}^{h} W(a_k, g_0) C_r(h - g_0) f_g(g_0) dg_0 + \int_{h}^{a_{k+1}} W(a_k, g_0) C_l(g_0 - h) f_g(g_0) dg_0 \\ &\quad + W(a_k, a_{k+1}) \int_{a_{k+1}}^{\infty} C_l(g_0 - h) f_g(g_0) dg_0 \end{aligned} \tag{7.27}$$

Using Leibnitz's rule and the fact that both $C_r(0)$ and $C_l(0)$ vanish at the origin,

$$\begin{aligned} M_k'(h) &= \int_{a_k}^{h} W(a_k, g_0) C_r'(h - g_0) f_g(g_0) dg_0 - \int_{h}^{a_{k+1}} W(a_k, g_0) C_l'(g_0 - h) f_g(g_0) dg_0 \\ &\quad - W(a_k, a_{k+1}) \int_{a_{k+1}}^{\infty} C_l'(g_0 - h) f_g(g_0) dg_0 \end{aligned} \tag{7.28}$$

Since $M_k'(a_k) < 0$, there is no relative minimum at a_k. Since both $C_r(0)$ and $C_l(0)$ are strictly convex, $M_k''(h) > 0$ for $a_k \leq h < a_{k+1}$ and there is either a unique minimum for some $h_1^*, a_k \leq h_1^* < a_{k+1}$, or there is a local minimum as h approaches a_{k+1}.

case (ii). $h \geq a_{k+1}$

$$\begin{aligned} M_k(h) &= \int_{a_k}^{a_{k+1}} W(a_k, g_0) C_r(h - g_0) f_g(g_0) dg_0 \\ &\quad + W(a_k, a_{k+1}) \left\{ \int_{a_{k+1}}^{h} C_r(h - g_0) f_g(g_0) dg_0 + \int_{h}^{\infty} C_l(g_0 - h) f_g(g_0) dg_0 \right\} \end{aligned} \tag{7.29}$$

$$\begin{aligned} M_k'(h) &= \int_{a_k}^{a_{k+1}} W(a_k, g_0) C_r'(h - g_0) f_g(g_0) dg_0 \\ &\quad + W(a_k, a_{k+1}) \left\{ \int_{a_{k+1}}^{h} C_r'(h - g_0) f_g(g_0) dg_0 - \int_{h}^{\infty} C_l'(g_0 - h) f_g(g_0) dg_0 \right\} \end{aligned} \tag{7.30}$$

Since $M_k''(h) > 0$ for $h \geq a_{k+1}$, there is a either a unique minimum for some $h_2^* \geq a_{k+1}$ or a local minimum at a_{k+1}.

In order to find the optimal solution, h^*, note that if $M_k'(a_{k+1}) > 0$, then $h_1^* < h_2^*$ and therefore $h^* = h_1^*$. If $M_k'(a_{k+1}) \leq 0$, then $h_1^* \geq h_2^*$ and $h^* = h_2^*$. The solution is unique, and the function $M_k(h)$ is unimodal under both cases (i) and (ii), so numerical search techniques may efficiently find the optimal solution if no exact solution can be derived.

Appendix 7B: Evaluating Expected Forecast Error Costs

In this appendix we derive algebraic expressions for $M_k(h)$, the expected forecast error accumulated from update epoch a_k to update epoch a_{k+1}. Assume that $w(t)$ is a constant between each potential update epoch, e.g. for all k, $w(t) = w_k$ for $a_k < t \leq a_{k+1}$. Let $L = a_{k+1} - a_k$, the length of time between update epochs. The definition of $M_k(h)$, equation 7.13, simplifies to:

$$\begin{aligned} M_k(h) &= \int_{a_k}^{a_{k+1}} (t - a_k) w_k C(t - h) f_g(t) dt \\ &\quad + L w_k \int_{a_{k+1}}^{\infty} C(t - h) f_g(t) dt \end{aligned} \tag{7.31}$$

for $k = 1 \ldots n - 1$ and

$$M_n(h) = L w_n \int_{a_n}^{\infty} C(t - h) f_g(t) dt \tag{7.32}$$

Also assume that the cost function is linear:

$$C(t - h) = \begin{cases} C_r |h - t| & \text{if } g_0 < h \\ C_l |t - h| & \text{if } g_0 \geq h. \end{cases} \tag{7.33}$$

With these cost and penalty functions, $M_k(h)$ can be expressed in terms of two integrals of the prior probability density function of take-off times. Define

$$I_{g1}(a, b) = \int_a^b t f_g(t) dt \tag{7.34}$$

$$I_{g2}(a, b) = \int_a^b t^2 f_g(t) dt. \tag{7.35}$$

Later we derive algebraic expressions for $=I_{g1}$ and I_{g2} for normal and gamma distributions. First, for $M_k(h)$ there are three cases corresponding to three ranges for the forecast h. The following expressions are derived from equation 7.31 and apply for $k = 1 \ldots n - 1$. Expressions for $k = n$, derived from equation 7.32, are similar.

case (i) $(h < a_k)$:

$$M_k(h) = w_k K_l \left[I_{g2}(a_k, a_{k+1}) - (a_k + h) I_{g1}(a_k, a_{k+1}) + a_k h (F_g(a_{k+1}) - F_g(a_k)) \right]$$

$$+Lw_kK_l\left[I_{g1}(a_{k+1},\infty)-h(1-F_g(a_{k+1}))\right] \tag{7.36}$$

case (ii) $(a_k \le h < a_{k+1})$:

$$\begin{aligned} M_k(h) &= w_kK_r\left[-I_{g2}(a_k,h)+(a_k+h)I_{g1}(a_k,h)-a_kh(F_g(h)-F_g(a_k))\right] \\ &\quad +w_kK_l\left[I_{g2}(h,a_{k+1})-(a_k+h)I_{g1}(h,a_{k+1})+a_kh(F_g(a_{k+1})-F_g(h))\right] \\ &\quad +Lw_kK_l\left[I_{g1}(a_{k+1},\infty)-h(1-F_g(a_{k+1}))\right] \end{aligned} \tag{7.37}$$

case (iii): $h \ge a_{k+1}$

$$\begin{aligned} M_k(h) &= w_kK_r\left[-I_{g2}(a_k,a_{k+1})+(a_k+h)I_{g1}(a_k,a_{k+1})-a_kh(F_g(a_{k+1})-F_g(a_k))\right] \\ &\quad +Lw_kK_r\left[h(F_g(h)-F_g(a_{k+1}))-I_{g1}(a_{k+1},h)\right] \\ &\quad +Lw_kK_l\left[I_{g1}(h,\infty)-h(1-F_g(h))\right] \end{aligned} \tag{7.38}$$

Finally, if the random variable g is distributed as a normal or gamma random variable, the integrals I_{g1} and I_{g2} can be expressed in terms of the distribution function and probability density function of g. First, let $g \sim N(\mu,\sigma^2)$. Let $f_g(t)$ and $F_g(t)$ be the probability density function and cumulative distribution function, respectively, of g. It can be shown that:

$$I_{g1}(c,d) = \sigma^2(f_g(c)-f_g(d))+\mu(F_g(d)-F_g(c)) \tag{7.39}$$

$$\begin{aligned} I_{g2}(c,d) &= \sigma^2\left[(c+\mu)f_g(c)-(d+\mu)f_g(d)\right] \\ &\quad +(\mu^2+\sigma^2)\left[(F_g(d)-F_g(c)\right]. \end{aligned} \tag{7.40}$$

Now let g have a gamma distribution with parameters a and b. Let $G(c;a,b)$ be the cumulative distribution function of a gamma with these parameters. Using the definition of the gamma distribution, it can be shown that:

$$I_{g1}(c,d) = ab\left[G(d;a+1,b)-G(c;a+1,b)\right] \tag{7.41}$$

$$I_{g2}(c,d) = b^2a(a+1)\left[G(d;a+2,b)-G(c;a+2,b)\right]. \tag{7.42}$$

Although these expressions may seem complex, they can be evaluated quite quickly on

the computer and are much more efficient than resorting to numerical integration for $M_k(h)$.

Chapter 8

Conclusions

The goal of this thesis is to generate accurate forecasts of aircraft take-off times from major domestic airports using information obtained in real-time by the FAA's air traffic management system. Towards this goal, we developed models for aircraft delays at the gate, flow models of airfield congestion, and models of individual aircraft movements from the gate to the departure runway. We also designed procedures for efficiently updating these forecasts, given the continuous flow of information to the system. The statistical models and procedures were formulated and tested with information that is either currently available to the FAA's Enhanced Traffic Management System (ETMS) or may be gathered in real-time from other sources within the FAA or from the carriers.

Besides developing statistical models, this thesis distinguished those sources of information that are most valuable for accurate take-off time forecasts. This information about the potential merits of various data sources may help to guide design decisions as the FAA upgrades and expands the data collection and forecasting components of the traffic management system.

This chapter contains three sections which summarize the thesis results, describe their implementation, and discuss their implications. Section 8.1 describes the primary models and empirical results. Section 8.2 makes recommendations for data gathering and model selection. Section 8.3 outlines areas for further research.

8.1 Primary Models and Results

Many air traffic management decisions are derived from projections of air traffic demand at fixes, sectors and airports. These projections are based on predictions of individual aircraft trajectories. Therefore, errors in trajectory forecasts produce errors in demand forecasts, leading to sub-optimal air traffic management decisions. Improving take-off time forecasts is of particular importance since uncertainty in take-off time is the primary source of forecast errors downstream. A simple model of Poisson arrivals at an airport demonstrated the relationship between take-off time forecast errors and errors in forecasts of arrival demand. The model found that that there are substantial demand forecast errors, given current levels of uncertainty in aircraft take-off times. In a numerical example, on average half the arrivals in a fifteen-minute period are surprises and half the predicted arrivals are no-shows.

The primary contributions of the thesis are dynamic, statistical models for take-off time prediction. Statistical analyses determined the primary explanatory variables for variations in ground transit times, the time between scheduled push-back and actual take-off. The models were constructed from these factors and were tested with additional historical data to determine their forecast accuracy. The models and empirical results are described in Sections 8.1.1 and 8.1.2, respectively.

8.1.1 Models for Take-off Times

The problem of take-off time prediction was divided into two components: predicting gate departure delay (delay at the gate after scheduled push-back) and predicting roll-out times (time from the gate to take-off). If these two quantities were strongly correlated, then a model sharing information between the two should also be developed. However, the analysis in Chapter 3 found that these two quantities are only weakly correlated in our data set. We therefore assume that the two are statistically independent and develop separate models for each. Take-off time forecasts are derived by adding predicted gate departure delays and roll-out times to the scheduled departure times from the gate.

Models for gate departure delay were proposed and tested in Chapter 4. Many of the longest gate departure delays were caused by the late arrival of a flight on its previous leg, i.e. delays of departures from Logan are often caused by delays at an upstream airport. Models built from other factors, such as weather and runway configuration, were much less

successful in predicting gate delays. Based on these observations, we constructed simple 'aircraft turn models' to link late arrivals with late departures. Since carriers have distinct patterns of turn times, we developed different gate delay models for each carrier.

For predicting roll-out times, we hypothesized that departure queues would play an important role, and in Chapter 5 we constructed aggregate flow models of aircraft departures in order to predict the number of aircraft rolling out. These predictions were then used as input to models which predict roll-out times of individual aircraft. While the aggregate flow models were originally developed to provide congestion estimates for individual roll-out time predictions, these models are also significant on their own as tools for monitoring airport capacity and predicting congestion.

In Chapter 6, we implemented variable selection and outlier detection procedures in order to develop static linear models of roll-out times. These models include as explanatory variables carrier identification, weather forecasts, runway configuration, and forecasts of airfield congestion. We also developed a set of dynamic models which respond more readily to the roll-out times of the most recent flights.

In Chapter 7 we developed procedures to efficiently incorporate real-time information into take-off time forecasts. Once an initial forecast has been produced, information about the aircraft's status continues to flow into the air traffic management system. The existing system monitors whether the aircraft has departed and updates the forecast when it expires, i.e. when the forecasted take-off time has passed and the plane has not yet departed. Our procedures determine optimal forecasts for each update, as well as the times at which updates should occur. The solutions are derived by minimizing a forecast error cost function which keeps track of forecast errors as they accumulate over time. A penalty is also assessed whenever a forecast update is made, to account for the cost of revising each forecast. Near-optimal forecasts and update schedules are generated using these procedures as well as simple heuristics.

8.1.2 Empirical Results

All of these models were tested with a data set assembled from a variety of sources. Data collection and sampling procedures are described in Chapter 3. In general, the models had mixed success in producing accurate forecasts. The gate delay models produced substantial improvements in forecast accuracy, and were successful in predicting many of

	% reduction, all flights	% reduction, actual GDD > 30 min.	% reduction, actual GDD > 60 min.
Aircraft Turn Model	50	56	57

Table 8.1: Percentage reduction in squared gate departure delay (GDD) forecast error achieved by use of the aircraft turn model instead of an aggregate mean. Results are shown for flights operated by 'Carrier B' at Logan Airport during the summer of 1991.

	% reduction, all flights	% reduction, actual roll-out > 30 min.	% reduction, actual roll-out > 60 min.
Static Linear Roll-out Model	17	11	0

Table 8.2: Percentage reduction in squared roll-out time forecast error achieved by use of the static linear model instead of an aggregate mean. Results are shown for roll-out time predictions made with a 3-hour forecast horizon for flights by the major carriers at Logan Airport during March and August, 1991.

the longest delays at the gate. The models produced a 40% decrease in the mean absolute error (MAE) when compared to a simple forecast of the overall mean gate departure delay. Table 8.1 shows that the model is also successful in predicting the longest delays. The table displays the percentage improvement in forecast accuracy (as measured by mean squared forecast error) over the forecast accuracy achieved by using an aggregate mean as a forecast. The table shows the results for one major carrier ('Carrier B'), and results for the other two carriers in our sample were similar. Over all flights by Carrier B, the mean squared forecast error from the aircraft turn model is half the mean squared error from the aggregate mean. The advantages of the turn model are even greater when the sample is restricted to the longest actual gate departure delays, those over 30 and 60 minutes. For all carriers, approximately half of the longest delays are left unexplained by the turn model. Delay codes recorded by the carriers indicate that these unexplained delays are likely to have been caused by mechanical difficulties, late passengers, or crew duties.

One important caveat for these turn models is that they assume we have perfect forecasts of arrival delays, since arrival delays are used to calculate subsequent departure delays after the aircraft turn. This is an optimistic assumption if we are producing forecasts three hours before the scheduled departures of the flights. However, predictions of arrival delays have been shown to be quite accurate once an aircraft has left the ground, and the turn

models developed in this thesis should significantly improve forecast accuracy for most flights one or two hours in advance. We also noted that the turn model will not accurately predict departure delays for flights which benefit from a 'swap,' e.g. when a late aircraft is exchanged for a flight which may be able to depart earlier. Model performance will be poorer at airports where swaps are more common, such as hubs.

The roll-out time models improved forecast accuracy over current methods, but were unsuccessful in anticipating the longest roll-out times. The static linear model driven by forecasts of airfield congestion (Table 6.2) produced a 24% reduction in squared forecast error over a short forecast horizon. A linear model with rough forecasts of departure demand (Table 6.1) produced the best results over a three-hour time horizon: a 17% reduction in the squared forecast error. This improvement did not apply to the longest roll-out times (Table 8.2). For actual roll-out times of over 30 minutes, the static linear model produces an 11% improvement. For flights with roll-out times greater than one hour, the predictions from the static linear model were no more accurate than the aggregate mean. These flights are delayed by unusual events which are not anticipated by the model.

For short forecast horizons, the dynamic linear models did not offer any advantages over the static linear models which included estimates of airfield congestion, possibly because temporary changes in airfield conditions were already incorporated into the congestion estimates. Nor did the dynamic linear models offer a clear advantage for forecasts over longer horizons.

Forecasts from the linear models were more accurate than a lookup table similar to the current system used in the ETMS. For example, the static linear model with the demand estimate produced a 15% improvement in MAE compared to the lookup table.

Estimates of airfield congestion were generated by the aggregate flow models, which monitored and predicted take-off capacity and the number of aircraft rolling out. Empirical tests found that the models predicted the number of aircraft on the airfield with a RMSE of 1.5 aircraft over a forecast horizon of 10 minutes, and a RMSE of 2.5 aircraft over a forecast horizon of one hour. However, these tests assumed that precise predictions of push-back times were available. Sensitivity analysis showed that RMSE over the shortest forecast horizon doubles if scheduled, rather than actual, push-back time are used in the predictions. Therefore, the flow model itself depends on accurate predictions of gate push-back times.

We also tested procedures for generating optimal take-off time forecasts given additional information about individual flights. For the ETMS, the information is limited to the fact that the aircraft had not yet taken off. We developed procedures for generating optimal forecasts and schedules to update those forecasts. The procedures were implemented for a variety of prior take-off time distributions, and near-optimal solutions were compared to solutions generated by simple heuristics. The differences between procedures were small when the standard deviation of the prior take-off time forecast is at a level which is typical in the existing air traffic system. For example, when the procedures were applied to a gamma distribution fitted to the residuals of a roll-out time model (with a standard deviation of 6.1 min.), the procedures differed by less than 5% in both forecast cost and expected number of updates. In general, the simple 'discrete conditional updating' procedure produced forecasts and update schedules that were close to the optimal.

The optimal updating procedures made greater gains when the prior variation in ground transit time was larger. For example, with a 40-minute prior standard deviation, the optimal update procedures reduced forecast costs by 25% while eliminating, on average, two forecast updates per flight when compared to the existing ETMS procedure. For large levels of prior uncertainty, the optimal updating procedures are successful in improving the accuracy of take-off time forecasts while diminishing the number of necessary updates.

8.2 Forecasting for Traffic Management: Implications and Recommendations

This section briefly describes the role of ETMS forecasts in selected traffic management procedures and makes recommendations for data collection and model implementation to improve take-off time forecast accuracy. Perhaps the most important implication of this thesis is that data collection and model selection are not independent. Given the existing information in the ETMS, small gains in forecast accuracy may be achieved by adopting models similar to the static linear models of Chapter 6. With the addition of flight connection information, the gains of the aircraft turn model of Chapter 4 may be realized. However, we found that even the most sophisticated forecasting procedures are unable to predict many of the longest aircraft delays given existing FAA information sources.

While this result may indicate that there is a significant level of irreducible uncertainty

in aircraft take-off times, it is also likely that information about many of the longest delays is in the hands of the carriers. Therefore, we first recommend better integration of data sources in the air traffic system by combining real-time FAA and carrier information. Relevant information from the carriers may be available far in advance of a flight's scheduled departure (as when an upstream delay will lead to a delay on a later flight), or may be available just before departure (as when a mechanical problem is discovered or an aircraft 'swap' will occur). Since this information was not included in this study, nor has it been generally available to the FAA, our models do not include real-time carrier information. Additional forecasting procedures should be developed if such information were available.

The following three sections elaborate on these points by describing information sources and appropriate models over long and short forecast horizons. Three forecast horizons are considered: flights pushing back from the gate in three hours, flights pushing back in 30 min., and flights that have, or should have, already pushed back from the gate. Finally, we see great potential for the development of an 'airport flow monitor,' which is based on the aggregate flow model of Chapter 5. This is presented in more detail in section 8.2.4.

8.2.1 Take-off Time Forecasts, Three Hours in Advance

Data from the ETMS currently serve as input into strategic traffic management programs, such as the national flow control programs coordinated by the Air Traffic Control System Command Center (ATCSCC) near Washington, DC. The primary strategic tool for flow control is the Ground Holding Program (GHP), which was described in more detail in Chapter 2. GHPs adjust projected arrival demand at congested airports to correspond with projected capacity by holding aircraft on the ground. Since the decision to hold an aircraft must be taken *before* the aircraft departs, GHPs are usually implemented hours in advance of the anticipated congestion. The GHP itself depends on forecasts of arrival demand and capacity at each airport of destination.

We have already mentioned the importance of carrier data in improving take-off time predictions, and the FAA, in their research plan, places a high priority on further collaboration and data exchange with the carriers [4]. The results of this thesis support the assertion that carrier delay data will substantially improve take-off time forecasts over long forecast horizons.

Given the existing information in the air traffic management system, and with the

addition of carrier gate schedules, an aircraft turn model would be valuable for predicting gate departure delays. In this thesis the most accurate predictions of roll-out times were generated by a static linear model, such as the model described in Table 6.1. This model was derived from a simple least-squares regression of roll-out time at Logan on factors such as carrier/runway combinations and flight duration. A particularly important component of the model is a measure of departure congestion: the number of scheduled departures close to the flight. Such a measure is easy to derive from existing ETMS data. Empirical tests demonstrate that this static model produces forecasts superior to those of the dynamic models, such as the lookup table and the dynamic linear model.

8.2.2 Take-off Time Forecasts, 30 minutes in Advance

There are a large number of traffic management procedures that rely on take-off time forecasts with relatively short forecast horizons. These include *departure metering*, *local ground delay programs*, and *airport initiatives*. Like a GHP, all of these procedures reduce demand downstream by holding aircraft on the ground. However, these traffic management procedures are limited to pairs of airports or particular regions. Lead-times for program implementation fall between fifteen minutes and an hour.

For these traffic management activities, forecasts have shorter horizons and are used to predict congestion at nearby sectors and fixes. Accurate gate departure delay information is again of primary importance. If a carrier knows that an aircraft will be delayed at the gate for an hour, the demand calculation should not assume an on-time departure. In the absence of more detailed carrier data, the aircraft turn model would again be valuable.

With a lead time of 30 minutes, as with the three-hour forecasts, our roll-out time models improved forecast accuracy using information that is currently available to the ETMS. In general, the same static linear model used for three-hour forecasts should suffice for this shorter forecast horizon. Slight improvements in forecast accuracy may be gained by implementing a model which includes a real-time estimate of airfield congestion. Such a model is displayed in Table 6.2, and the airfield congestion estimates N_k are generated by the aggregate flow models of Chapter 5.

However, no general mathematical model will be successful in predicting the longest roll-out times. Even when runway configuration, departure queues, and numerous other factors were taken into account, we were left with forecast errors which were, on average,

25% of the actual roll-out time for departures from Logan. Here again, carrier information may lead to further improvements.

8.2.3 Take-off Time Forecasts for Aircraft in Ground Transit

Once an aircraft is scheduled to push-back, we say that the aircraft is 'in ground transit.' For these flights we will already have generated a forecast for push-back and roll-out times, but we may have the opportunity to change the forecast as more information arrives. Take-off time forecasts for flights in ground transit are used to generate congestion alerts for sectors or fixes near the departure airport, and these alerts lead to tactical decisions to alleviate congestion. One example is the *departure sequencing* which occurs at airports in busy air traffic regions, such as the areas around New York or Los Angeles. Many airports may be sending aircraft to the same departure fix , so it is up to the terminal area air traffic facility (the TRACON) to coordinate departing aircraft so that many aircraft do not arrive simultaneously. Effective departure sequencing requires accurate take-off time predictions of aircraft that are close to departure.

At this time, the TRACONs which conduct departure sequencing do not rely on the ETMS for information. Each has its own information system which is linked to local airports. This is a good thing, for the ETMS provides little data about aircraft that are in ground transit. Accurate short-term take-off time forecasts must be dynamic and should be updated with information about the aircraft's trajectory. Forecasts may be updated at the scheduled departure time, when gate push-back occurs, when the aircraft reaches the departure queue, and when the aircraft is cleared for take-off. The ETMS reports none of these things, and the best dynamic forecasts must be generated from prior probabilistic information about the ground transit time, as described in Chapter 7. The only relevant information received by the ETMS is the actual take-off of the flight. Given existing information sources and levels of uncertainty, we recommend adoption of a discrete conditional updating heuristic. This would replace the ETMS's existing constant interval updating procedure .

Whatever gains an improved updating procedure may produce, the ETMS could improve its take-off time forecasts to a greater extent by collecting real-time gate push-back information (this information is already available, in real-time, to the carriers). Updating the take-off time forecast at push-back should lead to substantial improvements in forecast

accuracy. Suppose that a flight pushes back from the gate ten minutes after the scheduled departure time. From the experiments in Chapter 6, we know that the roll-out time forecast that would be produced has a standard deviation of 4 minutes. Without push-back information, however, all we know is that the aircraft is still in ground transit; it may be at the gate or rolling out. Suppose, as well, that we have a prior probability distribution of overall ground transit time. The best estimate without push-back information is the conditional mean of the overall ground transit time for this flight, conditioned on the fact that ten minutes have elapsed and the flight has not taken off. For a typical flight in our sample, the optimal conditional forecast with no push-back information, given that ten minutes have elapsed, is a 35-minute GTT (for a prediction of take-off in 25 minutes). The mean roll-out time is 18 minutes, so the variance of our 'conditional' take-off time forecast without push-back information is:

$$\begin{aligned} \text{Var(conditional forecast)} &= \text{Var(roll-out time)} + (\text{conditional forecast} - E(\text{roll-out time}))^2 \\ &= 4^2 + (25 - 18)^2 \\ &= 65 \text{ min.}^2 \end{aligned} \tag{8.1}$$

This forecast has a standard deviation of 8 minutes, which is double the standard deviation with push-back information. Push-back information will significantly improve forecast accuracy if forecasts are updated when the information arrives.

In general, the value of push-back information stems from the relatively large variance in gate delays, as compared to roll-out times. In our data set, the mean and standard deviation of gate delay was 5 and 14 min. while the those of roll-out time were 18 and 8 min., respectively. Information about gate push-backs allows us to eliminate the possibility of the largest gate delays and restricts the forecast to the smaller range of possible roll-out times.

However, accurate *forecasts* of gate push-backs produced by the carriers may be more valuable than real-time push-back information. Since the air traffic managers use the forecasts to make a series of decisions over time, the costs of forecast inaccuracy accumulate over time as well. In the spirit of the optimal forecast schedules in Chapter 7, early but slightly inaccurate information may offer more cumulative benefits than an accurate push-back signal that can only be used at the last moment.

8.2.4 Airport Monitoring with the Aggregate Flow Model

Besides improved forecasts, real-time data on gate push-back times would allow the ETMS to serve as a valuable tool for monitoring airport performance and evaluating the effectiveness of new equipment and procedures. Summary statistics, updated in real-time, would give traffic managers a quick 'snapshot' of airport performance in the recent past. A flow model similar to those developed in Chapter 5 would generate forecasts for the near future. Historical data stored from the model would allow researchers to investigate the effects of weather and traffic management innovations on airport capacity. A monitoring system of this type has already been developed to answer questions about capacity and delays in Frankfurt, Germany [3]. A similar system is being developed for Atlanta Hartsfield [2]. However, both of these systems now serve only for data collection and monitoring; neither contains a forecasting component.

Figure 8-1 shows a mock-up of a display for air traffic managers which was generated with Logan airport data from August 27, 1991. The top graph displays a running count of the number of departures from the airport over hour-long periods. The center graph displays the number of aircraft rolling out, and the bottom graph shows a running average roll-out time for recent departures. The forecasts shown by dotted lines in the graphs are derived from the aggregate flow model of Chapter 5. This display would allow managers in the tower and at remote sites to monitor and anticipate departure queues and changes in demand and capacity.

8.3 Directions for Further Research

The increasing availability of data in the air traffic system presents many opportunities for research that should lead to improvements in both the forecasting and control aspects of air traffic management. Potential research may be grouped under three general headings: encouraging data exchanges with the carriers, improving forecasts, and strengthening the links between forecasting and flow management. These are described in the following sections.

Encouraging Data Exchange

One result of this thesis is the demonstration of the importance of real-time delay infor-

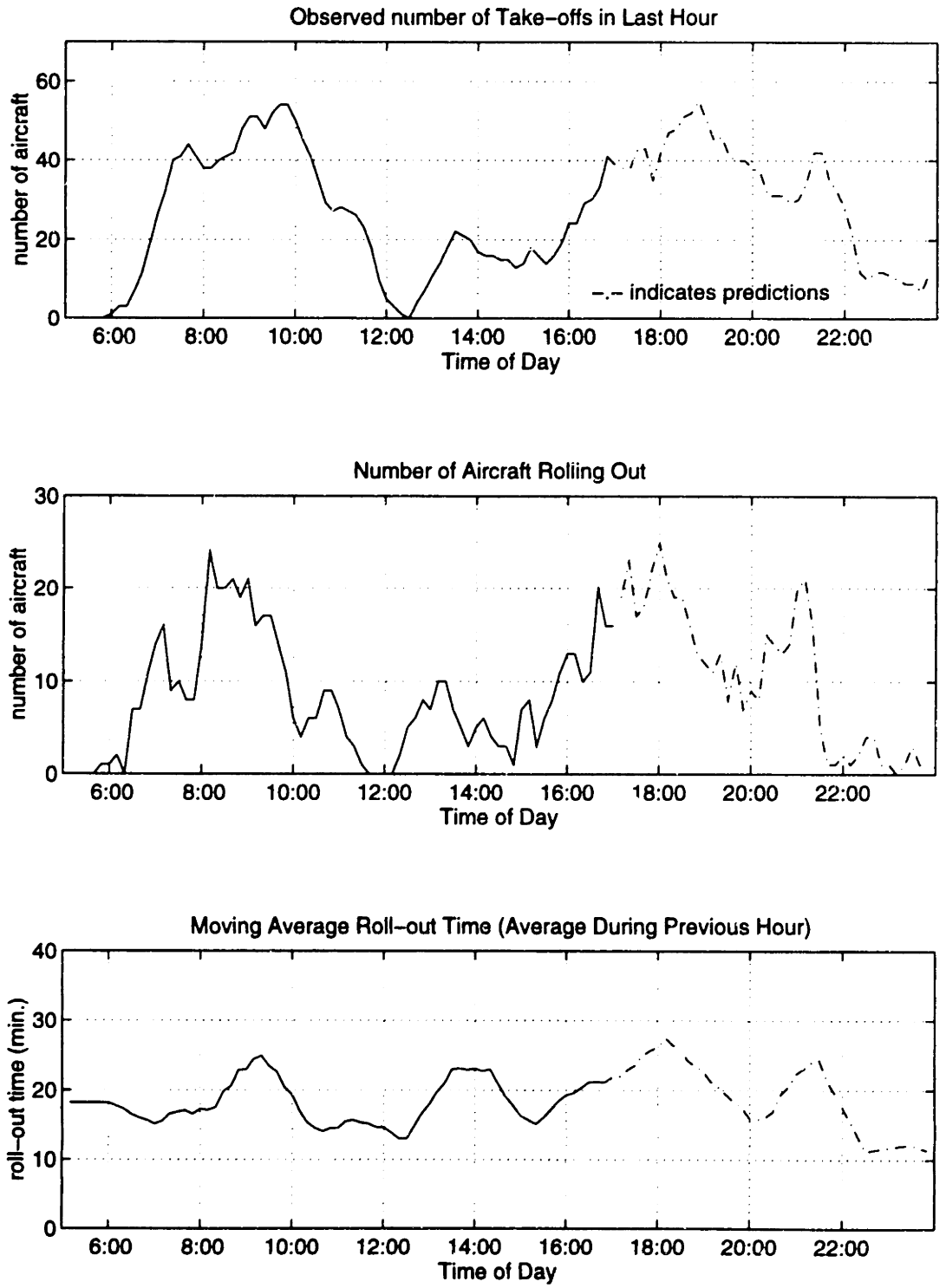


Figure 8-1: Mock-up of an airport departure flow monitor which keeps track of aircraft push-backs, take-offs, and roll-out times. Forecasts shown in the graphs are based on scheduled push-backs and capacity estimates from the aggregate flow model, using data at Logan Airport from August 27, 1991.

mation that is in the hands of the carriers. However, the carriers may be reluctant to share this information. They may fear that the information will be accessible to other carriers, placing them at a competitive disadvantage. They may also be worried that the FAA itself will not use the information fairly. For example, a carrier which shares cancellation information with the FAA must be assured that arrival slots which would have been used by the canceled flights will not be lost. The challenge to the FAA, therefore, is to design a data exchange system which encourages the carriers to contribute informative data, protects the carriers' competitive positions, and uses the data in an equitable manner.

Improving Forecasts for Air Traffic and Other Transportation Systems

If the models developed in this thesis are to be implemented, it will be important to adapt and test them at airports besides Logan. For example, in Chapter 3 we saw that the roll-out time profile at Atlanta Hartsfield was strikingly different from that of Logan, and the models must be adapted accordingly. In addition, the departure flow models of Chapter 5 may be extended to model the effects of multiple active runways, interactions between arrivals and departures, and 'feedback effects' when congestion affects subsequent demand at the airport.

Important methodological issues have been suggested by the work on optimal forecast updating in Chapter 7. These procedures may be generalized to find optimal forecasts, given a variety of information sources. Currently the procedures are limited to handling only the most rudimentary update information: the lack of an observed take-off. They may be extended to find optimal forecasts and update schedules, given information which changes the underlying probability distribution of take-off times.

In the long term, the air traffic management system will be linked with a surface movement tracking system [4]. This system will use ground radar and the global positioning system (GPS) to monitor the locations of all aircraft on the airfield. This would allow traffic managers to see each aircraft's position in the departure queue and would lead to many more opportunities to update forecasts. The data could also be used to locate bottlenecks on the airfield and guide aircraft to the least congested taxi-ways and runways. Procedures must be designed to efficiently handle this larger set of data.

The statistical models described in this thesis may be applied to other transportation

systems, and applications may be found in the burgeoning area of Intelligent Transportation Systems (ITS). For example, there is much interest in the use of real-time information systems to provide travel-time information to commuters. For example, one system which has been implemented by London Transport provides passengers at bus kiosks with forecasts of bus arrival times. This system has suffered from forecast inaccuracies similar to those that have plagued the ETMS [35]. The statistical and analytical techniques developed in this thesis may be applied to improve forecasts of bus, subway, and passenger car travel times.

Linking Forecasts with Flow Management

Research and development for the ETMS has focused on the technological challenge of gathering and disseminating enormous amounts of data across a nationwide network of air traffic control facilities. As a result, the ETMS has evolved into a storehouse of air traffic data which traffic managers may choose to use or ignore. Relatively little effort has been made to fit the ETMS into a framework which includes air traffic managers' decisions and the carriers' responses. The FAA seems to recognize this inadequacy and has made plans to develop "real-time decision making and analytical support tools for traffic flow management" [4]. Therefore, an important area of research is to quantify the costs of noisy data and forecast uncertainty. In terms of decision theory, we are curious about the value of perfect information. To find this we must first answer two difficult questions. What is the best traffic management strategy, given the data we have now? How much worse is this strategy than the best strategy with perfect knowledge of the future?

Finding the 'best' traffic management strategy is an enormously difficult task. We have seen that there are a wide variety of traffic management tasks, from a national GHP to the sequencing of departures on a single runway. Many of these procedures are interrelated, and some may be used to correct for inefficiencies in the others.

In addition, the almost continuous data stream from the ETMS opens up new areas for optimal control in the air traffic system. The information may be used as input to traffic management procedures that are more dynamic than those that have been considered. As new information about weather, capacity, aircraft delays and cancellations arrive, the optimal traffic management policy will change. This creates difficult decisions and trade-

offs. As was the case for the optimal forecast updating procedure, there may be a cost to changing a policy as well as gains from instituting a new one. Dynamic traffic management, efficient re-optimization, and inclusion of 'update costs' in traffic management strategies are rich areas for further research.

In this thesis, we have examined numerous models of aircraft departures from major airports, tested their efficacy, and examined their effects on air traffic management decisions. We hope that these models and the recommendations derived from them lead to improvements in the efficiency of the air traffic system while maintaining today's extraordinary level of safety.

Bibliography

[1] A. L. Breitler. *The Consolidated Operations and Delay Analysis System*, 1993. Presented at ORSA/TIMS Phoenix.

[2] T. R. Abrahamsen. Summary of the Atlanta Airport Resource Management Tool. Technical Report F045-M-050, MITRE-CAASD, Washington, DC, February 1994.

[3] Air Traffic Control Center, Frankfurt, Germany. *Flow Monitor*, 1993. Presented at a meeting on Air Traffic Control in the People's Republic of China.

[4] ASD-500. 1995 FAA Plan for Research, Engineering and Development. Technical report, Federal Aviation Administration, Washington, DC, December 1994.

[5] F. Proschan Barlow, R. E. and L. Hunter. Optimum Checking Procedures. *SIAM*, 11:1078–1095, 1963.

[6] J.M. Bates and C.W.J. Granger. The Combination of Forecasts. *Operations Research Quarterly*, 20(4):451 - 468, 1969.

[7] Gary Bishop. Oral Communication, December 1992. Gary Bishop is a regional traffic management specialist at the ARTCC in Nashua, New Hampshire.

[8] Alfred Blumstein. The Landing Capacity of a Runway. *Operations Research*, 7:752 - 763, 1959.

[9] A. M. Bonvik. Estimating the Lead Time Distribution of Priority Lots in a Semiconductor Factory. Working Paper OR 293-94, Operations Research Center, Massachusetts Institute of Technology, May 1994.

[10] L. M. Brown. Results of an Analysis of American Airlines Data. Technical Report WN 94W013, MITRE-CAASD, Washington, DC, April 1994.

[11] L. M. Brown. Oral Communication, May 1995. Lee Brown is a staff member at MITRE-CAASD in McLean, Virginia.

[12] L. M. Brown, V. A. Folmar, M. White, et al. Model Development Reflecting Uncertainties. Technical Report MTR 93W0000125, MITRE Corporation, McLean, Virginia, January 1994.

[13] E. A. Cherniavsky and J. H. Sinnott. Aircraft Itinerary Generator Documentation. Technical Report MTR-89W00100, MITRE Corporation, McLean, Virginia, June 1989.

[14] D. K. Sharma. Amendments to the On-Time Disclosure Rule. In *Federal Register*. Federal Aviation Administration, Washington, DC, September 1994.

[15] Federal Aviation Administration, Washington, DC. *Review of the FAA Research Engineering and Development Program: An Update*, January 1993.

[16] Federal Aviation Administration, Washington, DC. *FAA Strategic Plan, 1994*, 1994.

[17] Federal Aviation Administration, Washington, DC. *Terminal Area Forecasts for FY 1993 - 2010*, September 1994.

[18] Robert Feister. Oral Communication, January 1993. Robert Feister is a traffic management specialist at the Air Traffic Control System Command Center.

[19] Eugene P. Gilbo. Airport Capacity: Representation, Estimation, Optimization. *IEEE Transactions on Control Systems Technology*, 1(3):144–153, September 1993.

[20] Jesse Goranson. Looking for Trouble: How Well the FAA's Enhanced Traffic Management System Predicts Aircraft Congestion. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, August 1993.

[21] Gary Hofnagle. Oral Communication, February 1992. Gary Hofnagle is an engineer in the Logan Airport air traffic control tower.

[22] Richard Italiano. Oral Communication, May 1995. Richard Italiano is the Network Coordinator for American Airlines at Logan Airport, Boston, MA.

[23] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *The ASME Journal of Basic Engineering*, pages 35–45, March 1960.

[24] Bernard O. Koopman. Air-Terminal Queues Under Time-Dependent Conditions. *Operations Research*, 20:1089–1114, 1972.

[25] D. Luberoff et al. Bureaucracy to Boardroom: Crafting Air Traffic Control Inc. *Governing*, pages 55–58, January 1994.

[26] G. F. Newell. Airport Capacity and Delays. *Transportation Science*, 13(3):201–241, August 1979.

[27] Amedeo R. Odoni, Jean-Marc Rousseau, and Nigel H.M. Wilson. Models in Urban and Air Transportation. Unpublished Manuscript, 1994.

[28] G. Parmigian. On Optimal Screening Ages. *J. Am. Statist. Assoc.*, 88:622–628, 1993.

[29] Octavio Richetta and Amedeo R. Odoni. Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control. *Transportation Science*, 27(3):228 –238, August 1993.

[30] S. M. Ross. *Stochastic Processes.* John Wiley, New York, 1983.

[31] Mark Salanski and Others. FADE Investigation for Near Term Traffic Management Improvements: Interim Report #3. Technical Report FAA-AOR 200-94-001, Federal Aviation Administration, Washington, DC, April 1994.

[32] A. Seidmann and M. L. Smith. Due Date Assignment for Production Systems. *Management Science*, 27(5):571–581, 1981.

[33] A. Sen and M. Srivastava. *Regression Analysis: Theory, Methods and Applications.* Springer-Verlag, New York, 1990.

[34] B. Sengupta. Inspection Procedures when Failure Symptoms are Delayed. *Operations Research*, 28:768–776, 1980.

[35] Stephen T. Atkins. Passenger Information at Bus Stops (PIBS). Technical report, London Transport, London, England, January 1994.

[36] P. Stynes. Off Times vs. DZ Times. Technical Report Unisys/ATMS Project Memorandum, Volpe National Transportation Systems Center, Cambridge, Massachusetts, 1992.

[37] P. Stynes. Departure Queue Delay Studies. Technical Report Unisys/SAIC DTS-922, Volpe National Transportation Systems Center, Cambridge, Massachusetts, 1994.

[38] P. Stynes and T. Woytaszek. Basic Results of the DQD Data Analysis. Technical Report ETMS Project Memorandum, Volpe National Transportation Systems Center, Cambridge, Massachusetts, 1992.

[39] Volpe National Transportation Systems Center, U.S. Department of Transportation, Cambridge, Massachusetts. *Enhanced Traffic Management System Functional Description, Version 4.2*, VNTSC-DTS56-TMS-003 edition, January 1992.

[40] Peter B. Vranas, Dimitris J. Bertsimas, and Amedeo R. Odoni. The Multi-Airport Ground-Holding Problem in Air Traffic Control. *Operations Research*, 42(2):249–261, March-April 1994.

[41] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 1989.

[42] M. Zelen. Optimal Scheduling of Examinations for the Early Detection of Disease. *Biometrika*, 80:279–293, 1993.