Looking for Trouble: How Well the FAA' s Enhanced Traffic Management System Predicts Aircraft Congestion

by

Jesse Goranson

B.S. in Management Science, Massachusetts Institute of Technology (1992)

Submitted to the Sloan School of Management in partial fulfillment of the requirements for the degree of

Master of Science of Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1993

Author ______________________________
Sloan School of Management
August 30, 1993

Certified by ______________________________
Professor Arnold I. Barnett
Sloan School of Management

Accepted by ______________________________
Professor Thomas L. Magnanti
Codirector, Operations Research Center

Looking for Trouble: How Well the FAA's Enhanced Traffic Management System Predicts Aircraft Congestion

by

Jesse Goranson

Submitted the the Sloan School of Management in partial fulfillment of the requirements for the degree of Master of Science in Operations Research at the Massachusetts Institute of Technology. August 1993

ABSTRACT

This thesis tests the reliability of the FAA's Enhanced Traffic Management System, which uses various techniques to predict future aircraft congestion in the domestic air-traffic system. We describe the accuracy of the model in its predictions about individual aircraft as well as about aggregate congestion levels. We also consider how congestion might be defined for the purpose of calling "red alerts."

Thesis Advisor: Professor Arnold I. Barnett
Sloan School of Management

Acknowledgments

Many poor sods were sucked into helping me with this thesis, both directly and indirectly. Professor Barnett, of course, deserves much approbation for witticisms and questionable puns beyond the call of duty, deftly lightening what was a stressful time for me. More importantly, his clarity of thought in both verbal and statistical realms is legendary, and bittersweet, inasmuch as it usually leads to more work. I thank him for his guidance. Ken Howard, who I met at the Volpe Transportation Center, also deserves thanks. He spent more time with me than a sandaled graduate student usually rates, and deserves a raise. My roommate Stephan is as sincere and steadfast a roommate as anyone could ask for, and I thank him for his support through what has been truly an endless summer. I must also thank the Government, which through all its anfractous budget-making managed to do something good – pay for my education. (More specifically, I thank the Volpe National Transportation Systems Center and the FAA for their financial support under contract number DTRS–57–92–C–00054) And finally, I thank my parents, who are always there to remind me who and where I am.

Other people who deserve thanks, but I couldn't think of anything funny to say about, are: Larry McCabe, all the employees at Volpe Center that had the time to sit a spell with me, the students of the Operations Research Center at M.I.T., Alan Peyrat, Chris Couch, and my pledge brothers at Lambda Chi Alpha.

Table of Contents

List of Figures and Tables

Table **Page**

Figure

Chapter 1

Introduction

The Federal Aviation Administration is now immersed in a far-reaching project designed to increase the role of computers in air-traffic control. The computers are meant to assist in two ways, the first conceptually simple but extremely powerful, and the second representing a momentous step forward for air-traffic control. This computer system, tagged the Enhanced Traffic Management System (ETMS), collects weather data, aircraft scheduling data for aircraft on the ground, and position, altitude and velocity data for aircraft in the air. It then organizes the data into a flexible visual display that allows controllers to look at the current situation in a variety of ways. This first function collates data and extends the vision of controllers, allowing them to see the panorama of air traffic flow for the full country.

The second, more innovative function is predictive. The computer uses an assortment of interacting models to predict where every domestic aircraft will be at some future instant – for an average of about 6,000 aircraft in the air. The system uses these predictions to forecast above-ground congestion before it happens, and automatically notifies the air-traffic controllers monitoring the system by calling alerts specific to the areas, called sectors, requiring attention. The air-traffic controllers then consider amending the traffic flow to these sectors.

The ETMS is envisioned to do more than predict air traffic patterns. The FAA proposal includes developing an optimization package that looks at the ETMS-produced predictions and works out optimal routings for the whole system, keeping flow as safe and as steady as possible. However, the value of congestion alerts (let alone of any future optimization of routings) is compromised unless the predictions that underlie them are sufficiently accurate. The system can only realize its full potential if the predictions are sufficiently close enough to the mark; in addition, the alert system is only useful if the criterion for calling alerts accurately identifies situations that represent unacceptable levels of stress for controllers.

This thesis examines the accuracy of ETMS plane-specific predictions, and then investigates ways that those predictions might be improved. It also investigates how much this plane-specific forecast error reduces the accuracy of congestion alerts. (The current criterion for calling these alerts divides all future time into adjacent 15-minute periods; if the number of planes in a sector for any one of the fifteen minutes exceeds an FAA determined threshold, an alert is called.) We investigate alerts called under the current criterion – but also propose two alternate criteria for calling alerts that, while attractive, are not currently used. We examine whether these other criteria are more (or less) vulnerable to ETMS predictions errors than is the present standard.

Our findings are presented in detail in forthcoming chapters. After presenting some background material in Chapter two, we describe in Chapter 3 two earlier studies of ETMS accuracy, one by MITRE Corporation and the other by Ming-Cheng Chiang at the Flight Transportation Lab at M.I.T. In Chapters 4 and 5 we describe and calculate two plane-specific accuracy statistics, and investigate how the accuracy of predictions is reduced as those predictions are made further in advance. In Chapter 6 we investigate two ways to improve the system, and in Chapters 7 and 8 we examine the accuracy of congestion alerts as well as examine the value of two alternate criteria for calling congestion alerts. Our final chapter, Chapter 10, ties in the results of the various previous chapters, comments on possible causes of inaccuracy, and discusses some implications.

Our results portray a system that works well. While we found that predictions for planes on the ground are generally too optimistic, predictive precision was high for planes in the air. We also found that predictions are better for planes that are traveling through sectors at very high altitudes than at lower heights. One of our plane-specific results was remarkable: in general, the accuracy of predictions does not decrease as the predictions are made further in advance. As for alerts, we find that the system seems to err on the side of safety, meaning that the ETMS has more instances of calling alerts where none are needed than not calling alerts when they were needed. In fact, in a test based on real data, we found not one instance where ETMS did not call an alert when the number of planes actually entering a sector would have warranted one.

Thus pilots may rest safe knowing that the system is conservative, emphasizing safety at the expense of controller time. However, there is a trade-off: If the system "cries wolf" too often, controllers might get frustrated and trust the system less, making it a less useful tool.

Chapter 2

Background

The goal of the FAA's air traffic management system is to minimize delays and congestion while preserving safety and allowing users to schedule flights as their needs require. Because the demand for flights often outstrips the capacity of the system, especially at airports, the goal of air-traffic management often reduces to maximizing throughput.

The ETMS computers are designed to help the controllers analyze the entire system in real time. To achieve this, the FAA has segmented the development of the ETMS into five sequential phases. The first phase, nominally completed, is the development of the Aircraft Situational Display (ASD). This display reveals all current aircraft positions on a national scale, superimposed on maps of geographic boundaries and FAA air-traffic control facilities. The second phase uses the current aircraft data to forecast future demands on the system. Future phases will automatically suggest an array of possible routing solutions to traffic problems that arise, then evaluate which solutions are most desirable, and finally send the relevant parts of the optimal solutions to the specific air traffic control centers.

The air-traffic control system is organized as follows: The center of control lies in Washington, D.C., and is officially called the Air Traffic Control System Command Center (ATCSCC). ATCSCC keeps an eye on the whole system, and, if problems arise that threaten several large airports, ATCSCC runs an program that calculates how much to slow down all traffic connected to the affected airports in order to maximize throughput. This program is known to controllers as a "ground delay program", because many of its directives delay planes at their departure airports, thereby reducing the strain on the system.

The ATCSCC oversees 20 Air Route Traffic Control Centers (ARTCC's). These centers are distributed strategically about the United States, and are responsible for controlling problems that arise within their sphere of influence. These 20 spheres of influence cover the whole United States, and extend over onto the oceans and into Canada in some cases. The final tier of the air-traffic hierarchy is the airports themselves. The airports are concerned with very local problems in most cases – one of their main tasks is estimating how many aircraft can safely take-off and land given the weather conditions and other concerns particular to the moment. After determining these capacity restraints, the airports communicate with the ARTCC's, and often with other airports in an effort to space the aircraft properly.

Each ARTCC employs usually more than a hundred controllers who are constantly talking with the airline pilots, telling them what aircraft are around them, what lies ahead, and what changes in routing or altitude are safe and desirable for the whole system. The controllers maintain control over a small section of the ARTCC territory; these small sections are called sectors, and the entire territory of the ARTCC is divided into these subsections. These sectors are the smallest unit of control, and they can be characterized in a few ways. All sectors are three dimensional, and they are stacked, as well as abutting. Many sectors (but not all) fall into three general altitudes: low (0-24,000ft.), high (24000 to 34,000 ft.), and super-high (34,000 and above). Flights on longer routings are more likely to traverse these higher sectors than those on short hops, and some older planes and commuter planes are limited by altitude. Thus planes at different sector-levels may have different characteristics, whether those characteristics be type of plane, or length of route, or whether the plane is near the beginning, middle, or end of its journey.

The sectors also often have curious forms, dictated, in most cases, by the normal routings of the planes that fly through them. For example, the sectors around Atlanta are wedge shaped, and together form a spoke-like pattern – their main purpose is to shuttle flights in and out of Atlanta's airport in safe and timely manner. The sectors are also not usually carefully drawn out cubes or rectangles. They take irregular shapes to minimize the number of sectors a plane cuts through on its routings. Each time a plane moves from one sector to another there is a time-consuming hand-off between controllers, thus constructing the sectors to minimize the number of transitions saves time and energy.

Currently the ETMS does not model the irregular geometries of the sectors faithfully. The system designers have, however, begun a modeling process to remedy this problem that may well be concluded by the time this paper is completed. Our results, however, deal with a system in which the geometries are not properly modeled, and we point out a possible link between this modeling and our results in the final chapter of the thesis.

The ETMS makes demand predictions for sectors, fixes, and airports. Fixes are ground - based stations that emit signals telling a pilot where he is. They are contained within sectors, and thus from the demand data for sectors one could estimate the demand data for fixes. We concentrate only on sector demands, reasoning that sector demands and fix demands are so intricately linked that accuracy results pertaining to one can easily be applied to the other. We also avoid airports for the following technical reason: Often an aircraft will be asked to circle the destination airport several times before landing in order to make room for other traffic. These pre-arrival holding patterns introduce a volatility into ETMS landing predictions that has been described by Ming-Cheng Chiang, a student in M.I.T.'s Flight Transportation Laboratory. By dealing directly with sectors, we generally avoid this volatility, though some sectors that are very near airports may not entirely avoid it.

The first function of the ETMS, the Advanced Situational Display, allows the user to visually monitor what's happening with the system on a national scale in a variety of ways. The user can ask the system about any particular flight, and get its flight plan, its altitude, speed, and bearing. She could also ask, for instance, about all flights going into Chicago's O'Hare airport, about where lightning had struck in the last fifteen minutes, and about national patterns of precipitation. She could look at flights overlaid on maps of all low sectors, all high sectors, all states, or all ARTCC's. The ASD allows the controllers to do many more things, all intended to keep them well-informed.

Our concern, however, is with the second function of ETMS: the predictive capability. The ETMS allows the user to choose a sector, airport, or fix, and observe the predicted demand for each fifteen minute period ahead. Unless the user specifically asks for more far-reaching predictions, the system automatically displays the predicted demands for the four fifteen minute periods ahead (1 hour ahead total). The system also indicates what portion of the predicted demand is currently on the ground and what portion is currently in the air. The system does not allow the user to automatically list the specific aircraft that make up that prediction; he is merely presented with the aggregate demand data.

This chapter was meant to give a reader who is unfamiliar with the federal air-traffic control system enough information to understand the following analyses. The next chapter builds on this one by introducing the reader to some of the literature that precedes this study; it also illustrates how this thesis differs from and extends the work that has gone before.

Chapter 3

Literature Overview

There have been two studies on the accuracy of the ETMS predictions address issues related to our work. The first was authored by a team at MITRE, and the second constitutes the thesis of Ming -Cheng Chiang of the Flight Transportation Lab at M.I.T. We briefly summarize what these earlier researchers did in this chapter, and then explain how the current study complements and extends the previous investigations.

The MITRE study[1] used predictions of altitude, position, and direction that are made approximately three-minutes in advance of "now". For thousands of airborne flights, they compared these predicted positions with the actual positions three minutes later. To pin down the actual positions of the aircraft the MITRE people obtained raw data from the surveillance-based aircraft tracking system that is in place at the ARTCC's. The MITRE "three minute" prediction threshold was driven by the fact that, about every three minutes, the ETMS receives a position update that tells it where the plane actually is and thus renders any prediction of its present position obsolete. The researchers

[1] "A Preliminary Analysis of the Predicted Aircraft Position Accuracy of the Enhanced Traffic Management System Version 4.2", MITRE Co.

compared predictions just before they became obsolete to actual positions obtained from the tracking data.

The ETMS found that the positional discrepancies are reasonably small. In fact, the average X,Y, and Z discrepancies were 3.17, 2.76, and 0.38 nautical miles. We would expect that the discrepancies would be small, given that we are only predicting a few minutes in advance. What is surprising is that some (less than 0.1%) of the errors are very large . For instance, the highest discrepancies for each of the X, Y, and Z positional differences were 59.82, 73.52, and 6.51 nautical miles. In practical terms, this means that if the ETMS predicted a flight to be over Austin, Texas three minutes from now, in the worst case it could show up over San Antonio. The MITRE study determined that these errors are produced by either miscommunications between computer systems or potential problems with the ascent and descent modeling algorithms.

Ming-Cheng Chiang's study[2] considered some additional questions about ETMS accuracy. It concentrated on how skillfully the ETMS predicts airport arrival times for specific aircraft, where the predictions are made 10, 20, 30, 40, 50, and 60 minutes in advance of scheduled arrival time. The study calculated the average error and the standard deviation of that error for each of the six prediction times using data from hundreds of planes[3] . Though the analysis revealed the expected result that the errors seem to be normally distributed with mean zero, it could find no pattern for the error as a function of how far in advance the predictions are made; nor could it find a relationship between the standard deviation of the error and how far in advance the predictions were

[2] "A Study of Errors in Predicting Arrival Fix Times in Air Traffic Control", June 1993.

[3] Ming-Cheng Chiang's study used a sample of four airports of arrival: Denver, Orlando, Minneapolis, Phoenix.

made. One possible explanation is that many airplanes are placed in holding patterns at their arrival airports, and thus whether the system was predicting arrival times one minute in advance or an hour in advance, it was unable to anticipate these holding patterns. As a result, both predictions contained essentially a fixed amount of error. The result also means that although predictions far in advance as well as predictions less far in advance contain a certain amount of fixed error, the marginal error incurred by increased time between prediction and event is very small.

Cheng-Chiang went on to try and quantify whether predictive errors were concentrated in any one segment of the flight – in particular, the ascent and descent periods of the flight. It found no concentrations of error in those segments, a result that seemed to indicate that, though there was predictive uncertainty inherent in the system, it could not be attributed to problems with the ascent or descent models specifically.

This thesis expands on the work that has gone before in several ways. Controllers are not looking three minutes in advance as in the MITRE study – they are looking anywhere from one minute to hours down the road. Thus a characterization of predictive accuracy as the predictions are made further in advance is needed. Cheng-Chiang's study addresses this issue, but from the perspective of airport arrivals. Our thesis also addresses this issue, but from a sector perspective, thereby avoiding problems with holding patterns at arrival airports.

Our thesis also expands on the previous two studies by searching for ways to improve the system, based on previous characterizations of plane-specific errors. We push further by analyzing how these plane-specific errors compromise the system's ability to predict sector congestion, reasoning that, because the goal of the system is to reduce congestion, characterizing how effectively it achieves that goal is perhaps more useful overall than overmuch discussion of plane-specific accuracy.

Chapter 4

"Surprises" and "No-shows": Analysis

Our analysis of ETMS accuracy began with the estimation of two probabilities. First we estimated the probability that an airplane that had entered a sector had been predicted by the ETMS to do so. Put another way, we were estimating the number of "surprise" planes that arrive suddenly, without foreknowledge by the ETMS predictive apparatus. Our second estimate was of the probability that a plane actually enters a sector given that it was predicted to enter. In this case we were calculating the proportion of "no-shows", or planes predicted to enter the sector in question that just don't show up.

To estimate these probabilities we examined what happened to specific aircraft, and then aggregated their experiences to come up with a picture for the behavior of the "average" aircraft in the system. However, we also wanted to examine whether differences were evident between sectors at different altitudes. Towards that aim, we chose twelve sectors at random from the United States: 4 super-high, 4 high, and 4 low. We took lists of the nation's 426 low sectors, 274 high, and 80 super-high, and based on a randomization procedure involving the 1993 Boston White Pages, we obtained four sectors for each altitude level.[1]

ETMS keeps a running list of planes that are either predicted to enter a sector or have already entered a sector. The list is ordered chronologically, and thus the division between planes that have actually entered and those who are predicted to enter is determined by the time when the user looks at the system. For example if the user looks at the system at 6:00, all the planes that the system has recorded as entering the sector before 6:00 have actually entered, whereas all the planes in the list who are to enter the sector after 6:00 are predictions. Every minute the system generates a whole new list, and the division between actual and predicted planes increases one minute. For each of our twelve sectors we took a four hour excerpt of this list every five minutes for four hours, giving us 48 four-hour samples for each sector.

To illustrate, our first sample, at 3:05, contained all the planes that were predicted to be in a particular sector between 3:05 and 7:00 that evening. Thus our first four-hour block consisted entirely of ETMS predictions. Our 3:30 sample, however, contained actual information about when an aircraft entered a sector between 3:05 and 3:30, and predicted entry times for aircraft that were slated to arrive between 3:30 and 7:00. The list contained similar information for aircraft exit times. The 7:00 samples would consist entirely of actual information about what happened in the sector between 3:05 and 7:00. Thus if a plane actually entered the sector at 4:00 and left at 4:08, we could track the ETMS predictions for that plane by examining when it was predicted to enter and exit

[1] For this randomization procedure we used sequences of digits from telephone numbers at the top right-hand corner of consecutive pages to choose our sectors.

at 3:05, 3:10, all the way up to 5 minutes before it arrived at 4:00. (5 minutes because we took samples every five minutes, and thus the last prediction before 4:00 would be at 3:55) After 4:08, of course, the entry and exit times in our data set would be fixed, because the plane has come and gone, and no adjustments are made to its entry and exit times.

We were afraid that, if we only took samples for one day for each sector, our results might be susceptible to strange data quirks or bizarre storms. To allay that fear we took data for each of the sectors on a different day.

Our first statistic, the probability that a plane was previously predicted to arrive given that it did arrive in a sector, was calculated at follows: We chose a plane that we knew arrived in sector X by looking in the data set. For example, if a plane arrived at 4:00 and exited at 4:10, we would look at the 4:20 sample to pin down its entry and exit times. (As a rule we gave entry and exit times at least ten minutes to "settle". We found that occasionally it takes the system five to ten minutes after the fact to resolve exactly when a plane entered or exited a sector.) Then, we looked in the 3:50 sample for the aircraft. If the aircraft was predicted at 3:50 to enter the sector at any time between 3:50 and an hour after the plane exits the sector (where in this case the plane exits at 4:10), we consider the plane "predicted". To generalize, we chose a five minute period, say 4:00 to 4:04, and counted how many aircraft actually arrived during that period. We then examined the sample ten minutes before the beginning of that period (in this example 3:50), and counted how many of those planes that actually arrived in 4:00 - 4:04 were predicted to arrive in the time range 3:50 - 5:05. This "window of observation", as explained before, lasted from at least ten minutes before the actual entry time to at least an hour after the exit time. We performed this analysis for the 50 adjacent five-minute time periods from 3:00 to 6:30, for the twelve sectors, and over the two sampling periods, thereby obtaining data for over 1500 aircraft.

We also investigated how this probability changes as the predictions are made further in advance. After calculating the "surprise" statistic for ten minute predictions, we calculated the statistic for 20, 30, 40, 50, and 60-minute predictions. In each case our "window of observation" increased ten minutes so as to continue one hour beyond the expected arrival time, allowing us to continue to consider predictions that had the plane entering the sector later than expected. If we had not expanded the window for the increased prediction time, 60-minute predictions that had the aircraft coming in only a few minutes late would count as surprises, while 10-minute predictions that had the aircraft coming in 30 minutes late would not count as surprises. For example, without the increased "window of observation", flight TWA355 that was predicted at 4:00 to enter the sector at 5:00, but actually entered at 5:05, would be treated as a surprise even though the prediction was off only 5 minutes. We sought to avoid this possibility by expanding the window of observation.

Our second statistic investigated the probability that a plane actually entered a sector given that it was earlier predicted to enter. In this case we chose a hour time interval, say 3:00 to 4:00. We looked at our data sample for 3:00, and extracted all the aircraft that were predicted at 3:00 to enter the sector anywhere from 3:00 to 4:00. This method produces a variety of predictive "times-in-advance". In the current example we might have planes predicted to enter the sector at 3:05, 3:30, and 3:50. Thus we would have an array of "times-in-advance" including 5 minute, 30 minute, and 50 minute predictions. After obtaining these predictions for an array of individual planes, we evaluated whether those predictions held true by looking at the sample 1 hour beyond the previously chosen 1-hour interval. If the planes arrived in the sector anytime between the instant of prediction (in our example 3:00) and the time two hours ahead (in our example 5:00), we counted the planes as having shown up. Otherwise we labeled the planes "no-shows". In this statistic, unlike the previous statistic, the length of the "window of observation", is always fixed at 2 hours. Thus if an aircraft is predicted to enter the sector one-half hour after the beginning of the 2-hour interval of inspection (e.g. at 3:30), then the window of observation ranges from 1/2 hour before the prediction to 1 and 1/2 hours after (3:00 - 5:00). We performed this analysis for the three two-hour periods 3:00 to 5:00, 4:00 to 6:00, and 5:00 to 7:00 for the twelve sectors over the 2 sample periods, again obtaining data for over 1500 aircraft.

We define:

$\psi_z =$ P(Aircraft actually enters sector z / Aircraft is predicted to enter z)[2]

$\phi_{s,z} =$ P(Aircraft is predicted to enter sector z at time t_{-s} / Aircraft actually enters at time t_0)

Assuming that individual flights were essentially Bernoulli processes (e.g. show up/not show up), we used a Binomial assumption to calculate both ψ_z and $\phi_{s,z}$ from the plane specific data relevant to each statistic. This assumption allowed us to obtain maximum likelihood, unbiased estimates of the two probabilities. For ϕ_z this meant merely dividing the total number of planes that actually showed up by the number of these planes that were predicted to show

[2] This initial statistic is not dependent upon time-in-advance. Later, we examine the effect of predictive time-in-advance on the accuracy of predictions.

up for each of the individual "times-in-advance". We assumed that, once a plane failed to be predicted, it would be treated as a surprise even if it were correctly predicted to enter at some previous time. For example, a plane not expected to enter a sector twenty minutes in advance would be treated as a surprise even if it had been forecast to enter thirty minutes in advance (and had later been erroneously dropped from the arrival list). In practice, this convention meant that $\phi_{z,30}$ was the fraction of planes correctly forecast to show up one half hour before they actually appeared and also 20 minutes in advance as well as 10 minutes in advance.

For ψ_z we found the fraction of planes predicted to show up during "window of observation" that actually did so. To illustrate both statistics further, the probability that a particular plane was a complete surprise would be $1-\phi_{z,10}$ for a particular sector z. Similarly, the probability that a plane will not show up given that it was predicted to show is $1-\psi_z$.

We delved further into the behavior of ϕ by aggregating the results from sectors at the same altitude. We examined whether ϕ was appreciably different for super-high, high, and low sectors. We also investigated whether ϕ differed among commercial and non-commercial flights, our conviction being that commercial flights often have predictable schedules already entered into ETMS computers, while the flight plans from non-commercial flights are more tentative and, sometimes, misguided or lost . Thus fewer surprises might occur for commercial planes.

The statistic ψ_z revealed whether planes actually entered particular sectors as predicted, yet it said little about when they did so. Having calculated what percentage of the aircraft in our data were no-shows, we performed further calculations for those that showed up. Recall that we allowed the predictions to be off within a substantial "window of observation". Given that we had entry and exit-time predictions that varied in how far they were made in advance, we examined not only predictive accuracy, but inquired into how such accuracy changed as a function of time-in-advance.

We characterized predictive error in two ways. Instead of the immediately obvious actual entry time vs. predicted entry time, and actual vs. predicted exit time, we chose to compare the midpoint of the interval spent in the sector against the midpoint of the predicted interval. Because, arguably, the maximum stress a plane places on controllers occurs near the middle of its stay in a sector, we believe that error in estimating this midpoint is especially relevant to the controller. When planes are in the center of the sector, they are more likely to cross perpendicular flight paths (which should hardly be located near a sector's edge), and for that reason may especially need to be communicating with a controller. We also identify another type of error, that of "length of stay" in the sector. In this case "length of stay" error is actual exit time - actual entry time minus predicted exit time - predicted entry time. This statistic tells us whether the planes have stayed longer or shorter than originally anticipated, and thus whether the workload they imposed on controllers was greater or less than expected. The formulas for these quantities follow on the next page.

$$m = \text{ midpoint error } = \left[A_e + \frac{(A_x - A_e)}{2}\right] - \left[E_e + \frac{(E_x - E_e)}{2}\right]$$

$$l = \text{length - of - stay error } = (A_x - A_e) - (E_x - E_e)$$

where:

$$A_e = \text{ actual entry time}$$
$$A_x = \text{ actual exit time}$$
$$E_e = \text{ predicted entry time}$$
$$E_x = \text{ predicted exit time}$$

Note that the error in entry time and error in exit time may easily be deduced from these statistics. (i.e. given m and l, one can deduce E_e-A_e and E_x-A_x) For aircraft that did arrive given that they were predicted to, we wanted to get a sense of how far off the predictions were. Thus we calculated the statistics whose formulas are given above, and determined whether there were substantial differences across different altitudes by aggregating the results for each of the three altitude levels: low, high, and super-high. We also investigated how the errors behaved as predictions were made further in advance. First we categorized the observed m and l errors based on how far ahead the predictions associated with those errors were made. We then used regression analysis to determine whether those errors grew, declined, or remained unchanged as predictions were made further in advance. We performed the same regression using the average squared m and l rather than just m and l, in order to gain a sense of whether the "spread" of the error increased as predictions were made further in advance. We theorized that as predictions were made further in advance, the mean of the distribution of error would stay fixed at 0, while the variance of the distribution could be dependent on how far in advance the prediction was made.

Our final work with ψ_z involved investigating differences between the behavior of aircraft on the ground when predictions were made and aircraft in the air. Planes in the air would seem to proceed more predictably, and thus, much as ϕ might differ for commercial and non-commercial aircraft, ψ_z for planes on the ground may diverge from ψ_z for airborne planes.

The ETMS does not identify whether planes are in the air on a plane-by-plane basis, and although the ETMS does receive messages indicating whether an aircraft has taken off, these messages are difficult to get at en masse. To find out which of our aircraft were in the air we obtained routings and flight schedules from an on-line version of the OAG. Using that information, print-outs of the exact sector locations from the ASD, and information about average flight speeds of commuter planes and jets, we were able to deduce for each individual flight for which we had information whether it was in the air. We divided the aircraft into four categories:

- commercial flights that we knew were in the air at the time of prediction
- commercial flights on the ground at the time of prediction
- commercial flights for which we didn't know
- non-commercial flights

We then went back and calculated ψ_z for the four categories of aircraft, and performed tests of the statistical significance of differences among the three sector-altitude levels. More technical details about the analysis, as well as the results, follow in Chapter 5.

Chapter 5

"Surprises" and "No-shows": Conclusions

The results of our first analysis, in which we compute the proportion of planes that were anticipated visitors to a sector, follow in Table 1. As is evident from the table, the overall discrepancies between super-high, high, and low sectors are negligible. The very last column of the table represents probabilities calculated from the entire data set for each predictive "time-in-advance". Upon inspecting those probabilities, a simple rule of thumb arises: for each additional ten minutes we go back, about 2 out of 100 planes that enter the sector now had not been predicted to enter at all. If we observed that 50 planes entered a sector, we would expect two of those planes (4%) not to have been predicted to enter the sector as little as ten minutes before. Twenty minutes before, one additional plane would not have been predicted (2%); ten minutes earlier, another one falls out of the forecast, and so on. Thus there seems to be a fixed part of the error (4%), and a variable part, dependent on how far in advance the prediction is made (2% greater per 10 minutes back).

Table 1. Probability that a randomly chosen plane had been predicted to arrive in a sector given that it did, based on the type of sector and the time in advance.

Probability	Sector Altitude Groupings			
	Super High	Low	High	TOTAL
Time-in-Advance				
10 min.[1]	0.96	0.96	0.96	0.96
20 min.	0.94	0.94	0.92	0.93
30 min.	0.91	0.92	0.90	0.91
40 min.	0.90	0.90	0.89	0.90
50 min.	0.88	0.87	0.88	0.88
60 min.	0.87	0.86	0.86	0.86
Sample size:	513	477	449	1439

In Tables 2a and b we observe the proportions of planes predicted for commercial and non-commercial aircraft. Here stark differences emerge among the different types of sectors. In the high sectors, the commercial aircraft evince comparable patterns to the non-commercial aircraft. In the high sectors, however, as well as the low sectors, it seems that many non-commercial aircraft appear unexpectedly, while commercial planes are comparatively easy to predict. One might expect this result, given that commercial planes usually follow predictable schedules, but the numbers reveal striking differences. Where in the low sectors only 6 out of 100 commercial planes are surprises as far as an hour in advance, 35 out of 100 non-commercial flights were unanticipated an hour ahead. The differences in the low sectors persist even when the predictions are made

[1] Actually, between 10 and 14 minutes before, depending on when planes entered the sector. Thus 20 min. means 20-24 mins, exactly 10 minutes earlier than first prediction.

only ten minutes in advance: for commercial planes, only 1 out of 100 slips through the ETMS predictive net, a remarkable achievement. However, even ten minutes before, 13 out of 100 non-commercial planes are unanticipated.

Tables 2a and b: Probability that a randomly chosen plane had been predicted to arrive in a sector given that it did, based on the type of sector and the time in advance as well as the type of flight.

Table 2a: Commercial Aircraft

	Sector Altitude Groupings:			
	Super High	Low	High	TOTAL
Time in Advance				
10 min.	0.96	0.99	0.96	0.97
20 min.	0.95	0.97	0.91	0.95
30 min.	0.93	0.96	0.89	0.93
40 min.	0.92	0.95	0.88	0.92
50 min.	0.90	0.94	0.87	0.90
60 min.	0.90	0.94	0.85	0.90
Sample size:	411	350	333	1094

Table 2b: Non-commercial Aircraft

	Sector Altitude Groupings:			
	Super High	Low	High	TOTAL
Time in Advance				
10 min.	0.93	0.87	0.97	0.92
20 min.	0.89	0.83	0.95	0.89
30 min.	0.84	0.80	0.92	0.86
40 min.	0.80	0.76	0.91	0.83
50 min.	0.77	0.69	0.91	0.79
60 min.	0.76	0.65	0.87	0.76
Sample size:	102	127	116	345

Our "simple rule" of a loss of 2 planes out of 100 for each ten minute increase in the prediction interval averages quite different trends between commercial and non-commercial flights. While an OLS fit of the reduction in the probability over time for the combined data yields a slope of just above -2 (-0.019), the same OLS fit yields a slope of below -3 (-0.033) for non-commercial planes and a slope of -0.014 for commercial planes. The regression equations follow:

$$\phi_{combined} = 0.97 - 0.019T$$
$$(224) \quad (-16.8)$$
$$R^2 = .986$$

$$\phi_{\text{commercial}} = 0.97 - 0.014T$$
$$(166) \quad (-9.43)$$
$$R^2 = .956$$

$$\phi_{\text{non-commercial}} = 0.96 - 0.033T$$
$$(626) \quad (-84)$$
$$R^2 = .999$$

where T = predictive time - in - advance

In plainer terms, the ETMS loses about three non-commercial planes per ten minute increase in predictive interval while losing only about one-and-a-half commercial planes in the same interval. Figure 1 provides a graphical comparison.

Returning to Tables 2a and b, we note the disparity between commercial and non-commercial results for the low sectors. The predictions are excellent for commercial aircraft, horrid for non-commercial, and yet the average of the two is near the commercial/non-commercial averages for high and super-high sectors shown in Table 1. Predictions concerning non-commercial aircraft in low sectors are the worst of the whole set of results, while those predictions for commercial aircraft are the best – the two results cancel.

These discrepancies could arise because commercial flights tend to be very similar from day to day, following the same predictable routes and changing flight plans seldomly. Non-commercial planes, however, including military jets and corporate aircraft, have erratic schedules, and are therefore hard to predict. This non-commercial population of flights seems to be an subset where the ETMS does not do so well. In the end, however, the poor predictions concerning non-commercial planes are camouflaged by excellent predictions concerning the whereabouts of commercial flights.

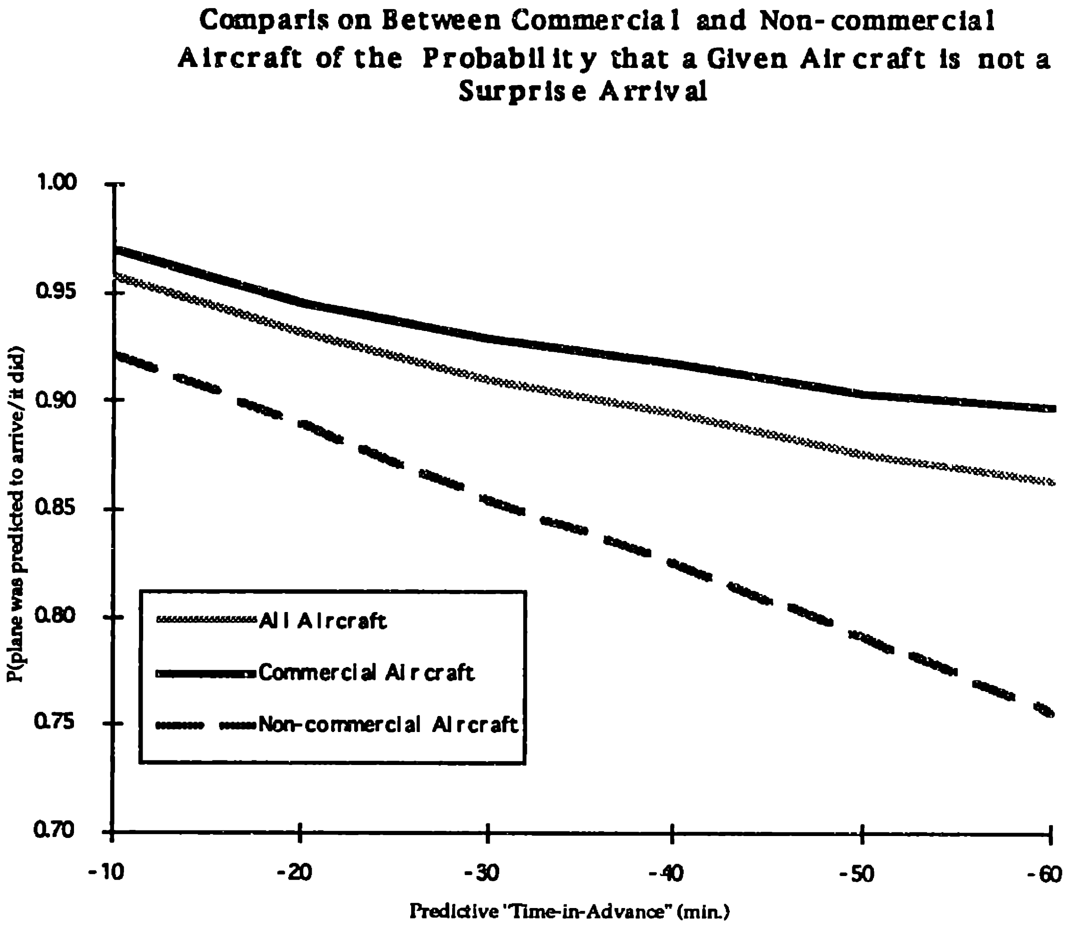


Figure 1. Graphical comparison of trend lines of ϕ as predictions are made further in advance. Trend lines represent differences between commercial and non-commercial flight data.

To test whether the differences in predictive accuracy between commercial planes and non-commercial planes were statistically significant, we performed a Binomial test of significance, using conditioned probabilities. (conditional probabilities because the events are not independent; for example we use P[see plane 20 minutes before/plane seen 10 minutes before] rather than P[see plane ten minute before]) The probability of the observed pattern, wherein the ETMS was better at predicting commercial planes than non-commercial ones six times out of six, was 0.016 (.5^6) under the null hypothesis of no difference in behavior between the two subsets. Thus the null hypothesis is clearly rejected under a two tailed test at the 0.05 level of significance.

Table 3 on the following page communicates the results of our ψ computations: the proportion of planes that enter a sector z within the aforementioned "window of observation" given that the planes were predicted to enter the sector. Those planes that did not show up though they were predicted to were deemed "no-shows". ψ is calculated for our four groups of aircraft: and for times in advance ranging from 1 min. to 60 min.

The flights in the air tend to do consistently better at avoiding no-show status, yet the super-high sectors seem to fall behind in this regard. In comparison to the previous results, the system seems to be better at avoiding surprises than identifying planes that are either exceptionally late, take different routes, or perhaps do not show up in the predicted sector for other reasons. This result is a bit surprising, because one would expect planes that are rerouted from one sector might show up as surprises in another, yet the probability of "no-show" is consistently higher than the probability of "surprise", as one can quickly surmise from a glance back at Tables 1 and 2.

Table 3. Probability that a flight actually arrives in a sector given that it was predicted to arrive, for different types of flight, different altitudes, and different status at time of prediction (in-air/on ground). Predictions range from 1 to 60 min. in advance.

	Status at time	Sector Altitudes	Sample Sizes
	of prediction	Super-high Sectors	
Commercial:	*In-flight*	0.83	225
	On Ground	0.80	197
	Unknown	0.75	135
Non-Commercial:		0.64	167
Overall rate:		0.76	
		Low Sectors	
Commercial:	*In-flight*	0.91	120
	On Ground	0.84	281
	Unknown	0.80	123
Non-Commercial:		0.64	191
Overall rate:		0.79	
		High Sectors	
Commercial:	*In-flight*	0.92	131
	On Ground	0.72	119
	Unknown	0.72	174
Non-Commercial:		0.77	194
Overall rate:		0.78	

Chi square tests of the discrepancies between sectors revealed significant differences between super-high, high, and low sectors for each category of planes (save the "unknown" category). Overall rates among the three types of sectors are similar. However, we conclude that within the four groupings of aircraft, differences arise. Super-high sectors tend to have the highest rate of "no-shows" for aircraft in the air, almost 10 aircraft per 100 flights higher than either of the other altitudes (83% for super-high vs. 91% and 92% for low and high sectors respectively). It is possible that, because planes tend to fly very close to the high/super-high border, they may often descend into the high sectors, especially when avoiding turbulence. Indeed, in Tables 2a and b we see that the commercial planes in high sectors have the highest probability of any of the sectors of showing up unanticipated by ETMS.[2] (the probability of surprise is 1- the probability of no surprise; the probability of no surprise is the statistic recorded in Tables 1 and 2) This phenomenon, where super-high flights "slip" into high sectors, might conceivably account for the higher rate of error both in the "no-show" statistic for super-high sectors and the "surprise" statistic for high sectors.

For aircraft that were not "no-shows", and did in fact enter the sector as predicted, we calculated midpoint and "length-of-stay" errors as described above in Chapter 4. (Recall that these predictions were made 1 to 60 minutes in advance.) The average midpoint error for all flights, including both commercial and non-commercial, and those on the ground and in the air, was +5.16 minutes, meaning that, on average, planes tended to be about five minutes late at the

[2] Changes in altitude require ATC approval – it is not the case that a change in altitude will catch the controller of the sector in which the change is made by surprise, yet it might mean that the new routing will carry the plane through sectors it would not otherwise have flown through. Thus a plane may be unanticipated ten minutes beforehand, but it will never suddenly spring upon a controller.

middle of their visits to the sector. The average "length-of-stay" error was +0.42 minutes, meaning that on average, planes tended take slightly longer to traverse the sectors than anticipated. For aircraft in the air, the respective averages were +0.68 and +0.57. Our estimates of those quantities for aircraft on the ground are 6.52, and 0.38. As expected, the in-air flights had substantially smaller errors than the combined data. The average length-of-stay errors, however, were very similar between the two groups.

We next investigated whether the sector altitude levels revealed differing patterns of prediction error. Figures 2 through 5 indicate what we found. We divided the 60 minute range of predictions into six ten-minute prediction periods, and took averages of the midpoint error for each range. In the first figure (Figure 2), midpoint errors were consistently worst for the <u>high</u> sectors in the analysis containing all aircraft. Low sectors did consistently better than high sectors, but consistently worse than super-high, which had the least amount of midpoint error in every one of the six time-period divisions. Figure 2, then, reveals that flights are on average late, not early; this result is not surprising given that many of the aircraft in the data used to create this chart were on the ground at the time of prediction. The ETMS uses a very optimistic adjustment algorithm for planes on the ground that are still on the ground when their scheduled departure time has come and gone. If a plane has failed to leave at (say) the predicted time of 4:10 PM, the ETMS "bumps" the plane's estimated departure time to 4:15, as well as all predictions dependent on the original departure time of 4:10. If the aircraft is still on the ground when the new, revised departure time arrives, the system bumps all predictions back another five minutes. Thus predictions associated with aircraft on the ground obviously reflect "best case" departure times given the information at hand; thus the associated optimism might be contributing to the lateness evidenced in Figure 2.

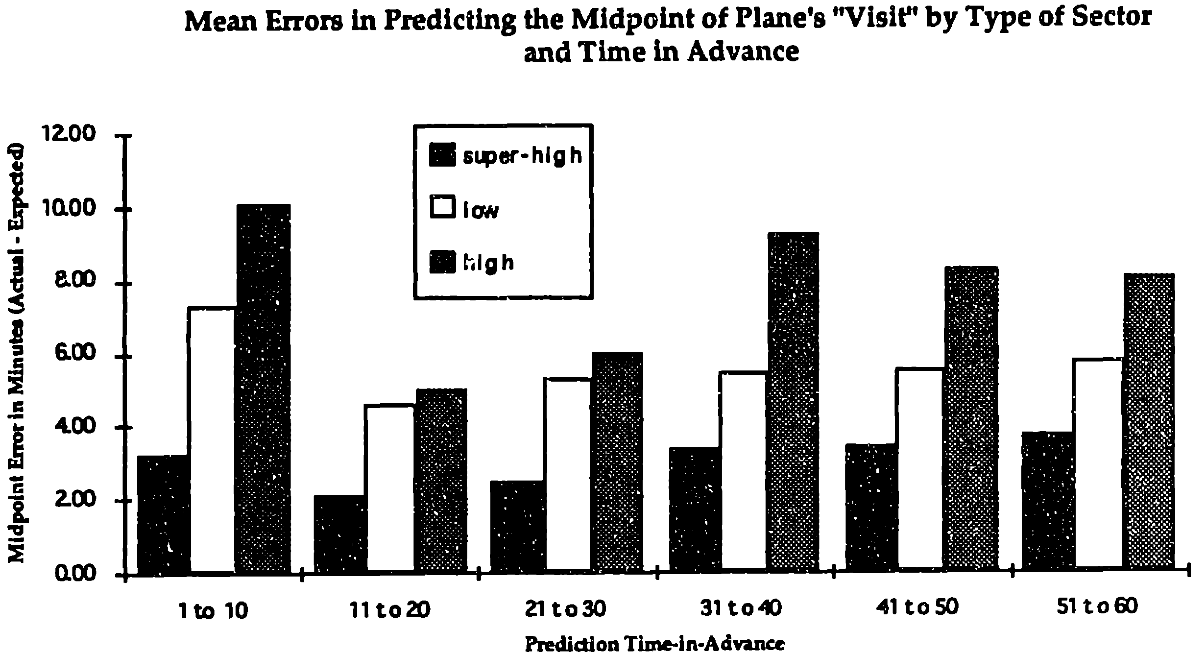


Figure 2. Bar Chart indicating consistent midpoint tardiness as a function of predictive time in advance, where the predictive time in advance is aggregated into 6 10 minute periods. The bar chart also illustrates relative levels of tardiness for the three classifications of sector altitude levels.

Figure 2 tells us two things: 1) that aircraft tend to be consistently late arriving at the middle of their path through the sector, 2) that high sectors tend to have the highest average errors, low sectors have the second highest, and super-high sectors have comparatively lower average errors.

The pattern revealed in Figure 3, where only commercial planes in the air are included in the midpoint error statistic, supports our theory that predictions for grounded planes tend to be overly optimistic. The pattern of lateness all but disappears. No clear trends are visible, but, again, even with planes that are in the air, high sectors tend to have the highest absolute level of error, while super-high sectors tend to be only a few seconds off. Overall, the average magnitude of these midpoint errors is very small, the largest average error is about a minute-and-a-half; even planes traveling 600 miles an hour can travel only ten miles in a minute, and many planes in our data set do not travel that quickly – some commuter planes travel less than half that fast. Even as far as an hour ahead of time, the system tends to be not more than a minute off for these in-flight planes, and thus the average distance between where a plane is and where it was predicted to be is on average small.

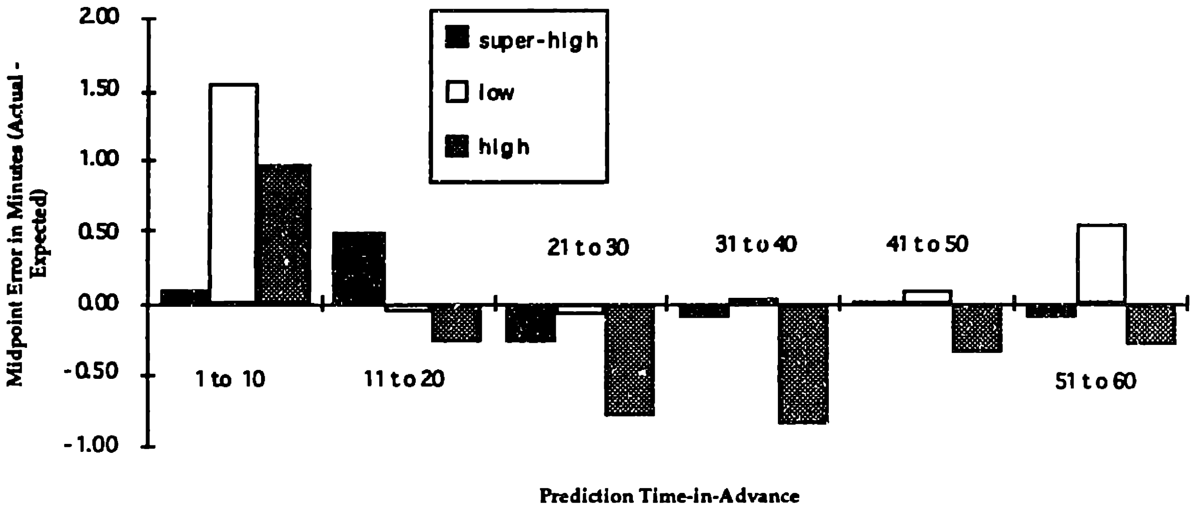

Midpoint Error Comparison Across 3 Sector Altitude Levels and Six 10-Minute Aggregations of Predictive Time-In-Advance, for In-Flight Aircraft Only

Figure 3. Bar Chart indicating midpoint error as a function of predictive time in advance, where the predictive time in advance is aggregated into 6 10 minute periods. The figure illustrates relative levels of error for the three classifications of sector altitude levels; the chart is based solely on in-flight data.

The length-of-stay error for all aircraft, as shown in Figure 4, reveals that aircraft consistently remain longer than anticipated in the sectors, though as the chart reveals, the average excesses are below one minute (the mean is below thirty seconds). Perhaps aircraft tend to stay longer than anticipated because the expected velocities of the aircraft are higher than the actual velocities. But Figure 3, showing nearly zero-mean normal patterns of midpoint error for flights in the air, does not support this hypothesis.

Figure 4 below illustrates a small but positive average length-of-stay error for low and high sectors and for aircraft in the air, where we might expect a zero mean length-of-stay error. The fact that planes tend to be in sectors longer than anticipated, even when the predictions are made when the flights are in the air, reveais what might be an area of improvement for the system. Again, though, super-high sectors seem to have comparatively little length-of-stay error, indicating that the system does quite well in predicting where flights in those sectors are going to be, even as far as an hour ahead.

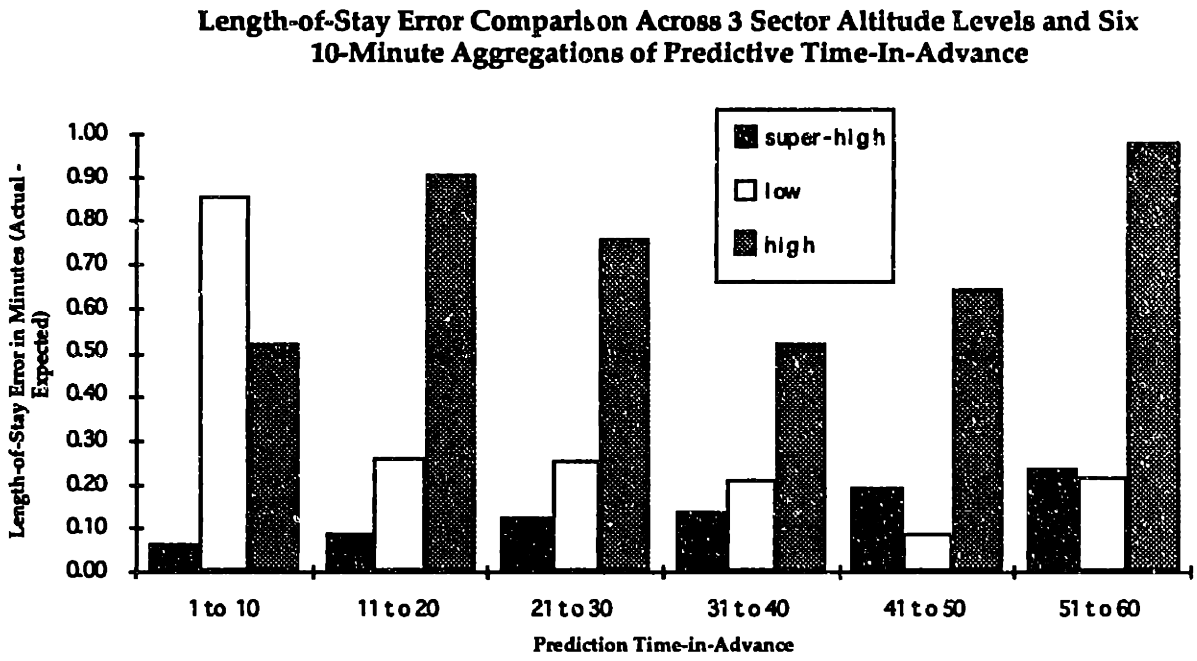


Figure 4. Bar Chart indicating length-of-stay error as a function of predictive time in advance, where the predictive time in advance is aggregated into 6 10 minute periods. The figure illustrates relative levels of error for the three classifications of sector altitude levels.

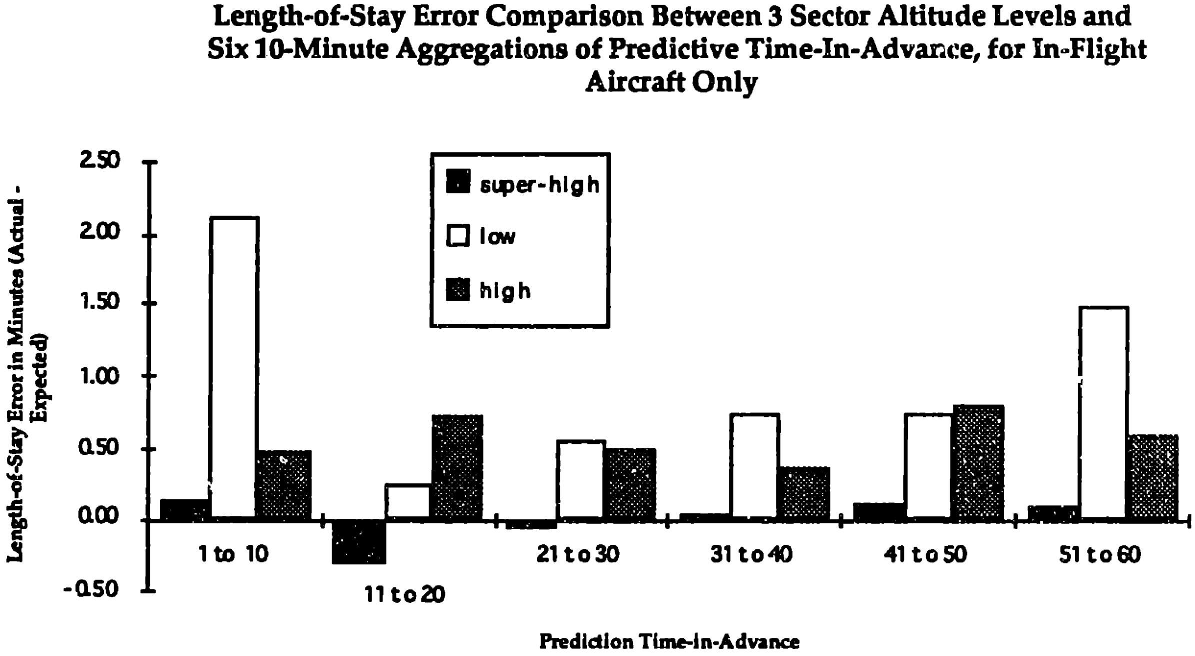


Figure 5. Bar Chart indicating length-of-stay error as a function of predictive time in advance, where the predictive time in advance is aggregated into 6 10 minute periods. The figure illustrates relative levels of error for the three classifications of sector altitude levels; the chart is based solely on in-flight data.

Our next step was to examine whether the average errors got systematically more positive or more negative as a function of how far in advance predictions were made. In other words, were 60-minute predictions on average less accurate than 10-minute predictions? This is a question of absolute magnitude, not overall average. We would expect that errors would bob about a mean 0, and, although the variance of the error might grow as the predictions are made further in advance, the mean should not change. OLS regressions using average midpoint error or average length-of-stay error as dependent variables and time as the independent variable produced intercepts, coefficients, R-squared values, and t-statistics as follows. Results are calculated using the whole data set as well as calculated using in-flight data only; T is time-in-advance in minutes.

Regressions Using Entire Data Set

$$m_{\text{combined}} = 5.48 - 0.00T$$
$$(6.19) \quad (-0.14)$$
$$R^2 = .0003$$

$$l_{\text{combined}} = 0.17 - 0.01T$$
$$(0.87) \quad (2.06)$$
$$R^2 = .098$$

Regressions Using In-Flight Data Only

$$m_{\text{in-flight}} = 0.77 - 0.01T$$
$$(1.65) \quad (-0.73)$$
$$R^2 = .013$$

$$l_{\text{in-flight}} = 0.06 - 0.02T$$
$$(0.13) \quad (1.47)$$
$$R^2 = .054$$

Based on the results of these regressions, we see that there is little cause to believe that the average midpoint errors are affected by how far in advance predictions are made. The slope of error vs. time was 0.00 for all flights (!) and a mere 0.01 for in-air flights. The regression on *l* using the complete data set reveals a slight relationship between time-in-advance and length-of-stay, but that relationship disappears in the in-flight data result. Both t-statistics for the intercepts of the length-of-stay and midpoint error regressions on the in-flight data are below 2 in absolute value, meaning that 0 is within the confidence interval for the intercepts. Thus the hypothesis that the errors for the in-flight data are varying about a mean of 0 cannot be rejected.

Given that theory supports a zero-mean hypothesis and our regressions do not refute it for in-air data, the variance at each of the sixty data points (predictions made one minute in advance through those made 60 minutes in advance), can be estimated by the average squared error of the collection of flights grouped into each predictive "slot". Thus for the 32 minute prediction "slot" (i.e. forecasts 32 minutes in advance), there might be three planes with average midpoint errors of 3 minutes, -2.5 minutes, -4 minutes, and 7 minutes.

Their average squared deviation around 0 would be about 20 (i.e. [9 + 6.25 + 16 + 49]/4).

While the mean predicted error might be expected not to vary with time, the mean-squared error could well do so. We regressed these average squared errors against time to discover if, indeed, the average error, be it positive or negative, increased linearly as the predictions were made further in advance. Our theory was that, even though the errors might be zero mean, their variance might increase linearly with time, in accordance with the well-known pattern in "random walk" theory. We also reasoned that if the magnitude of the error increased linearly with time, its squared error could increase with the square of time in the following manner: Suppose the ETMS is making predictions based on an estimated speed that is slower than the actual speed. If the ETMS predicts when the plane will reach a point 600 miles from its present position, its error will be twice as large as that for a point 300 miles away. Thus the squared error would be four times as large. We performed OLS regressions of squared error on both time-in-advance and the square of time-in-advance, using in-flight data for both midpoint and length-of-stay error.

General Equation:

$$e^2 = \alpha + \beta t + \gamma t^2$$

$e \equiv$ average error (whether it be midpoint error or length - of - stay error)
$\alpha \equiv$ fixed amount of squared error (intercept)
$\beta \equiv$ amount of squared error varying with time
$\gamma \equiv$ amount of error varying with time squared

$$m = 1.42 + 0.00T - 0.012T^2$$
$$(2.13)\ (0.12)\quad (\text{-}0.21)$$
$$R^2 = .005$$

$$l = 1.88 + 0.003T - 0.012T^2$$
$$(2.74)\ (2.72)\quad (\text{-}2.27)$$
$$R^2 = .20$$

$m \equiv$ midpoint error
$l \equiv$ length - of - stay error

The variance of midpoint error remains relatively constant, as shown in Figure 10. However, the length-of-stay regression reveals a significant relationship for both t and t^2. The coefficient for the square of predictive time-in-advance is negative, signifying that as predictions are made further in advance, the average squared error decreases. We performed two additional regressions of average squared error on time and time-squared separately: neither produced significant results. That fact, together with the relatively thin margin of significance and the anti-theoretic results led us to believe that the slight significance might be an artifact of the data – a curve-fitting result rather than a statistically meaningful one.

The results of the regressions convey that there is no time dependence. This is certainly not what we expect, for it is saying that predictions get no better as they are made closer to the time of the actual event. In Ming-Cheng Chiang's thesis the same result prevailed, but they were examining predictions of airport arrival times, not sector arrival times. They surmised that because aircraft were often put in holding patterns when they reached the airport of arrival, predictions made right before they were handed off to the airport's controllers (who initiated the holds) were just as bad as those made an hour before. However, our result is even more surprising, given that our sectors are encountered near the beginning, middle and end of aircraft routings, and thus are generally not affected by airport holding patterns.

In summary, we have done the following in this chapter:

- Discovered the following rule of thumb: for each additional ten minutes we go back, about 2 out of 100 planes that enter the sector now had not been predicted to enter at all, except for the first ten minutes, at which 4 out of 100 planes fail to be predicted.
- Concluded that commercial planes, perhaps because of the regularity of their routings, appear in a sector unanticipated by the system less often than non-commercial planes. This effect is especially pronounced in the low sectors.
- Concluded that predictions concerning aircraft in the air are much more accurate than predictions for those on the ground, and quantified that disparity.
- Estimated average midpoint and length-of-stay errors for flights in the air as well as flights on the ground, and found that flights on the ground tend to be late, perhaps due to optimism in the ETMS's modeling of departure times.

- Discovered that there is no relationship between either the average error or the average squared error and how far in advance the prediction associated with that error is made.

In the final section we discuss in more detail the implications of these results for the system as a whole. In the next section, however, we build upon the definitions of m and l in an effort to develop procedures that will improve the accuracy of predictions.

Chapter 6

Potential Improvements

Having explored how well the ETMS is doing at predicting if and when planes reach sectors as well as how long they stay there, we proceeded by investigating some ways that ETMS accuracy might be enhanced. We focused on two particular areas for potential improvement:

1) Exploring the relationship between entry time error and length-of-stay error, where entry-time error was calculated as the actual time a flight entered a sector minus the predicted time.

2) Attempting to establish whether errors in one time period gave us any clue about errors in the period immediately following, thereby allowing us to improve the accuracy of the system by incorporating that first-period information in the second-period estimates.

In our earlier analyses we concentrated on midpoint and length-of-stay errors. Here, however, we study entry-time error as a predictor of length-of-stay error. The main reason for the change is that almost by definition the two types of error are correlated. Suppose that the ETMS correctly predicts an aircraft's moment of entry into a sector, but incorrectly predicts how long it will stay in the sector. Then the midpoint error will be half the length-of-stay error, and of the same sign. Of course, we do not even know midpoint error until the plane has exited the sector, so using that error as a predictor of duration error is a hopeless venture from the outset.

Thus, in practical terms, we wanted to know if when the system revealed that the entry time was different from expected, we could make a better guess of how long the aircraft would stay in the sector. We chose the OLS regression technique because it would not only give us a sense of whether length-of-stay error was related to entry-time error through t-statistics and R-squared statistics, but it also would give us an idea of how much they were related by producing slope coefficients. Moreover, in simple bivariate regressions, the coefficient of correlation can be derived by taking the square root of R-squared, providing us with another well known measure of their relationship.

As in previous analyses, we concentrated on planes that were already in the air, reasoning that regressions based upon this data would be less likely to be influenced by extreme data points. Our fear was that unusual errors might exert power on slope and intercept estimates way out of proportion to their practical weight, thereby obscuring what might be interesting patterns. In many individual cases ground holds created entry-time errors of as much as 70 minutes, whereas most of the errors were below 10 minutes. The largest error of the in-flight data, however, was 17 minutes, and that measurement was one of 3 out of over 500 that was greater than ten minutes. Thus a certain amount of homogeneity was present in this latter set, a homogeneity that is often friendly to regression analysis.

We performed two regressions, one of the absolute value of length-of-stay error on the corresponding absolute value of entry-time, and one simply of length-of-stay on entry-time. The first regression would reveal whether the magnitude of the independent variable said anything about the magnitude of the dependent variable, regardless of sign. The second regression adds to that information by revealing whether errors of similar sign tend to follow one another – a result that, if significant, could allow the system to adjust its length-of-stay predictions based on entry-time error.

The results of these regressions follow. The equations, with their t-statistics, show that the absolute errors are related in a statistically significant manner; in particular, there is an average length-of-stay error equal to 1.23 minutes even when there is no entry-time error, and for each additional five minutes of entry time error there is approximately an additional minute of length-of-stay error. The t-statistic here is 2.95, illustrating that a relationship certainly exists, though not of overwhelming significance. The results allow us to predict, given an absolute entry-time error of five minutes, that the absolute length-of-stay error would be along the order of two minutes and five seconds in absolute value (0.17 * 5 + 1.23 = 2.08 minutes).

Absolute Length of Stay Error Regressed on
Absolute Entry Time Error

$$|l| = 1.23 + 0.17|m|$$
$$(6.63)\ (2.95)$$
$$R^2 = .025$$

Length of Stay Error Regressed on
Entry Time Error

$$l = 0.69 - 0.07m$$
$$(4.03)(-1.30)$$
$$R^2 = .005$$

However, this regression is preliminary, because it does not consider the sign of the length-of-stay error. Thus we turn to the second regression, which relates the two errors rather than their absolute values. The significance of the first regression fades away in the second. The best-fit estimate tells us that for each 15 minutes the plane comes in later than expected (giving us positive entry-time error), the length-of-stay error goes down by one minute, and, conversely, for each 15-minute period the plane comes in early, the length-of-stay error increases by one minute. (-0.07 minutes = 4.2 seconds; 4.2 seconds * 15 is approximately 1 minute).

The t-statistic of this second regression indicates that the slope is not significantly different from zero. Thus, while we are able to predict absolute length-of-stay errors based on absolute entry-time errors, we are not sure whether being late implies a longer or shorter than usual length-of-stay. Even if the result was significant, the slope is so small that, as a practical matter, the result is less than useful. It takes fifteen minutes of entry-time error to "produce" one minute of length-of-stay error, yet only one of the over 500 entry-time errors was over fifteen.

In the next analysis we explored whether errors in one period were related to the errors in the period following. To test the hypothesis we reasoned that predictive error in one direction of movement (e.g. westbound) might 1) be different, or even opposite, to predictive error in the opposite direction, and 2) that the difference in error might carry over from one period to the next. Specifically, if the way the model considered wind speed in its predictions was off for a certain day and a particular part of the country, we might be able to first characterize the error specific to a particular direction, say east, and then incorporate that characterization into subsequent easterly predictions, thereby making those subsequent predictions more accurate.

We determined the bearing of each flight in our in-air data set in the following way: We examined the routing of the aircraft, and then measured whether which of the following vectors was most appropriate to describe the aircraft's path: northbound, southbound, westbound, or eastbound. For example flight AAL55 might travel from Buffalo to Chicago's O'Hare field to the Dallas/Forth Worth airport. The sector at which we are examining midpoint and length-of-stay error is on the path from O'Hare to Dallas, so we concern ourselves only with that leg of the route, and it is most certainly southbound. We dropped flights for which it was not clear which vector they were following.

We also dropped seven of the twelve sectors from our analysis because they lacked enough directional data to produce meaningful results. For the remaining sectors, we maintained the division between the first day's sample and the second day's sample; our aim was to examine relationships between errors in the first half of our data set and the second half, not between days. Then for each of the 10 samples (2 for each of 5 sectors), we divided the data into flights traveling in opposite directions. For each sample we had only one pair directions with enough data for the exercise (i.e. N/S or E/W).

As an example, for sector ZJX50 we might have ten flights traveling northbound and eleven flights traveling southbound. We would divide those flights into a first period and a second period by taking the first half as the first period and the second half as the second period, breaking uneven divisions by sending the extra flight to the first period. Thus in this example we would have five northbound flights in the first north period and five in the second period. Similarly, we would have six southbound flights in the first south period and five in the second period.

Then we would calculate the average midpoint error for flights in the first period traveling (say) easterly. We would do the same for the westerly flights, and then repeat the calculations for the second period. If eastbound flights in the first period were, on average, later (larger positive midpoint errors) than the westbound flights from the same period, we would investigate whether that relative lateness of eastbound flights carried through to the second period. If eastbound flights tended to be earlier, we would determine whether they were earlier in the second period as well.

If there were in reality a relationship between the errors in subsequent periods, then whether direction one's average error was greater than direction two's in the first period should have no bearing on which is greater in the second. For the ten days we made this comparison for midpoint error, and found that eight of the ten directional comparisons yielded the same outcome in both periods (e.g. eastbound delays greater than westbound ones). Nine of the ten yielded consistent patterns for length-of-stay error. The days that did not are indicated in boldface in Table 4 below. We would expect, given no relationship between period one and period two that five of the ten comparisons would be consistent – under a Binomial assumption, the one-tailed p-value for the midpoint error result is .054, meaning that roughly 1 out of 20 times the observed deviation from the expected 5 out of 10 consistent comparisons would arise merely by chance. The p-value for the for the length-of-stay error is .011.

Table 4. Comparison of average midpoint error and length-of-stay error between two adjacent time periods for two opposite flight directions.

			Midpoint Error		**Length-of-Stay Error**	
sector	day	period	*Direction 1*	*Direction 2*	*Direction 1*	*Direction 2*
ZJX50	Day 1	*Period 1*	**-1.18**	**0.8**	-1.38	-0.8
		Period 2	**-0.21**	**-1**	-1.28	-0.5
	Day 2	*Period 1*	2.5	3.67	**0.6**	**-1.33**
		Period 2	-1.5	0.625	**-1.4**	**0.25**
ZJX67	Day 1	*Period 1*	**-1**	**-0.6**	0.67	-0.8
		Period 2	**0.25**	**-1.38**	-0.5	-1.25
	Day 2	*Period 1*	-0.38	-0.75	-0.75	0
		Period 2	0.5	0.38	-0.33	0.75
ZOB98A	Day 1	*Period 1*	-1.83	-2.5	-1	-0.33
		Period 2	-0.33	-1	0	1
	Day 2	*Period 1*	1.67	0.5	-0.67	0.2
		Period 2	-0.5	-1	-0.33	0.5
ZMA40	Day 1	*Period 1*	-0.21	1.62	-0.42	1.25
		Period 2	-0.83	5	-0.67	6.5
	Day 2	*Period 1*	1.75	-0.17	-0.25	1.67
		Period 2	0.21	-1	0.14	0.67
ZMP13	Day 1	*Period 1*	0	-1	0	0
		Period 2	3.16	-0.75	2.33	0
	Day 2	*Period 1*	2.2	-3.6	0.4	-0.4
		Period 2	3.3	-0.9	1	-0.2

Thus our calculations suggest some relationship between period one and period two error, particularly in the length-of-stay error calculations. This result could arise from problems in the way wind-speed is incorporated into aircraft speed predictions, but more extensive data is needed to pursue this relationship further.

We had calculated previously an average midpoint error (about +7 seconds) for all in-flight aircraft, and thus for each aircraft in the second period, regardless of direction, we subtracted off this "average total error", squared the result, and summed the squares. This first sum-of-squares statistic was a measure of how widely the errors varied from this overall average. In other words, if in each case we "guessed" that the error would be merely the average error of the whole system, this statistic gave us an aggregate measure of how far off we would be However, we wanted to calculate a similar sum-of-squares where our "guesses" were based on the average error in each direction from the period before. Thus, if the average error for eastbound planes was 1 in the first period, we would make 1 our guess for how late each eastbound plane would be in the second period. We would use a different guess, of course, based on the first period westbound average. Then, as before, we subtracted the "guess" from the actual error, squared the difference, and then summed the result. To test whether the second "guessing" scheme was any better than the first, we merely looked at which sum-of-squares was lower. The results of this analysis appear in Table 5 below.

Table 5: Comparison of sum-of-squared error for midpoint predictions and an adjusted sum-of-square error that takes into account error patterns from the immediately previous time period.

	Sample 1		**Sample 2**	
	Sum of Squared Error	*Sum of Squared Adjusted Error*	*Sum of Squared Error*	*Sum of Squared Adjusted Error*
Sectors				
ZJX50	47.2	61.0	68.0	166.6
ZJX67	52.8	**49.5**	48.9	55.5
ZOB98A	15.8	23.9	10.2	27.1
ZMA40	361.6	**319.4**	42.7	56.8
ZMP13	52.8	**52.3**	179.3	**166.0**

Our adjusted sum-of-square statistic does not do appreciably better than the unadjusted sum-of-square statistic. In only four of the ten days does the "improved guess" do better than an unadulterated guess. This result could reflect the dependence of the sum-of-squares statistic on a few extreme measurements; it does suggest, however, that there are limited benefits to using earlier errors from planes traveling in a given direction as a guide to errors for future planes from that direction.

Chapter 7

The Accuracy of Alerts

Thus far we have examined issues pertaining to plane-specific data, yet the ETMS aggregates these data and uses them for the specific purpose of calling alerts. Here we discuss three alternative criteria for issuing alerts and, in Chapter 8, consider whether the plane-specific data are accurate enough to support effective use of the ETMS warning system.

Recall from Chapter 2 that for sectors, the ETMS demand is measured as the peak number of aircraft in any one minute over a fifteen-minute period. Alerts are called when this demand exceeds a sector-specific threshold. It should be noted, however, that different situations could lead to the same maximum statistic. For example, a fifteen-minute period with five planes in each of the individual fifteen minutes would have the same maximum demand as one in which five planes were present in the first minute and none thereafter (i.e. the maximum number of planes over the fifteen minutes is five in each case). Because this insensitivity may reveal potential flaws with the maximum criterion as a measure of the level of demand on controllers, we also examine two other criteria in the following analyses.

The maximum criterion, by definition, assumes that a controller is most stressed when he has to deal with many planes at once for a single minute. However, perhaps a more stressful situation occurs when a controller has to deal with many planes over a long period. Taking the average number of planes-per-minute would reflect this increased stress by reflecting the number of "plane-minutes" that a controller will have to deal with over a fifteen minute period. (The former is the latter divided by 15) Perhaps these "plane-minutes" are a better measure of controller workload, as follows: For each minute that a plane is in the sector, a controller has to keep track of its position and talk it through the sector. As the number of plane-minutes over a fifteen minute period keep piling higher and higher, the controller is working harder, and the situation becomes more stressful. Because total plane-minutes over a quarter hour is simply the average number of planes then in the sector times fifteen, using the average number of planes over fifteen minutes may be more useful than considering only the 15-minute maximum.

However, the average statistic has its own problems. The average can smooth over very high values by tempering them with lower values over the fifteen-minute period. Thus if the number of planes suddenly doubled in a sector, the average would be pulled up, but perhaps not enough to convey properly how stressful such a situation would be to a controller. An average squared statistic would, however, take a high-variance situation like the one described into account. The average-square statistic is calculated by squaring the demand estimates for each minute of the fifteen-minute period, and then averaging them together. To use the same measure as for the other two criteria – number of planes – we take the square root of the average squared. Note that this value will be different than the average. The average square statistic would treat high-variance situations as particularly stressful. This criteria would make sense if, for instance, controlling five planes a minute over fifteen minutes was less stressful than controlling an average of five but rapid oscillations between two and eight. Through a simple manipulation of the well-known formula for variance, we see that the average-squared, or the second moment of a population, is equal to the variance plus the square of the mean: $s^2 = \mu^2 + \sigma^2$ Thus the average squared statistic counts both a large number of planes and much variance about the mean as stressful properties of demand.

We therefore considered three criteria for measuring stress/workload over a quarter hour:

- the maximum
- the average
- the square root of the average squared

For these three criteria we first calculated demand predictions as follows: Recall that for each of the 12 sectors we had four hours of data for each of two days. For each of these 96 hours of data (12 x 4 x 2), we determined the predicted demand level for the third quarter of the hour (30 min. to 45 min.). Then, by looking at the ETMS data for the end of the hour, we determined the true demand levels for that third quarter. This procedure gave us 96 pairs of predicted and true demand levels for the three criteria, where the predictions were made 30 min. in advance. We then graphed predicted versus actual demand for each criterion.

Figure 6

Plot of True 15-Minute Root Average Squared versus Predicted 15-Minute Root Average Squared for 96 Data Points from 12 Sectors

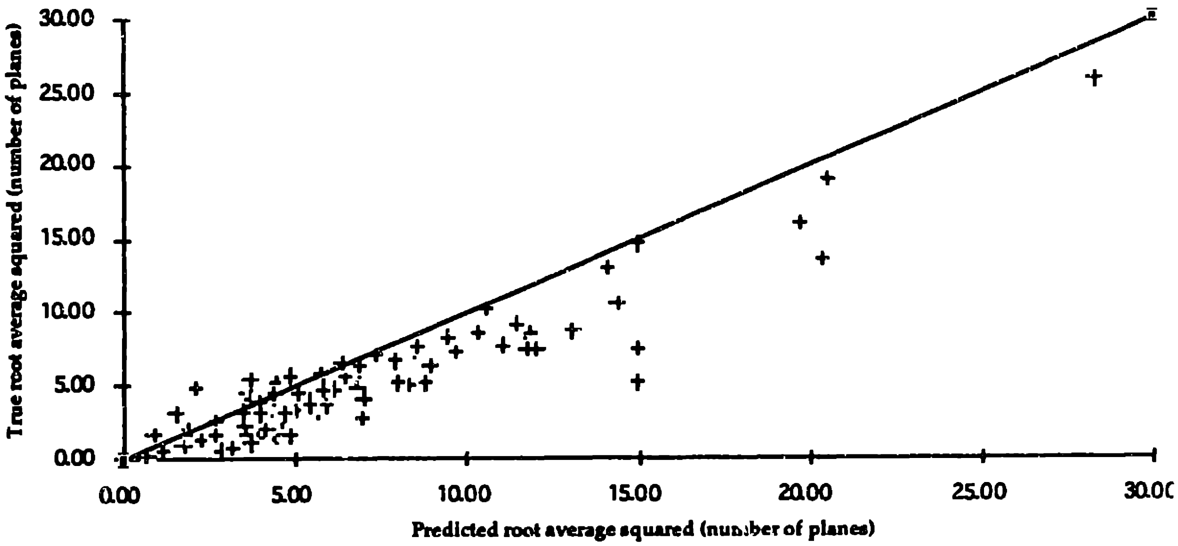


Figure 7

Plot of True 15-Minute Average versus Predicted 15-Minute average for 96 Data Points from 12 Sectors

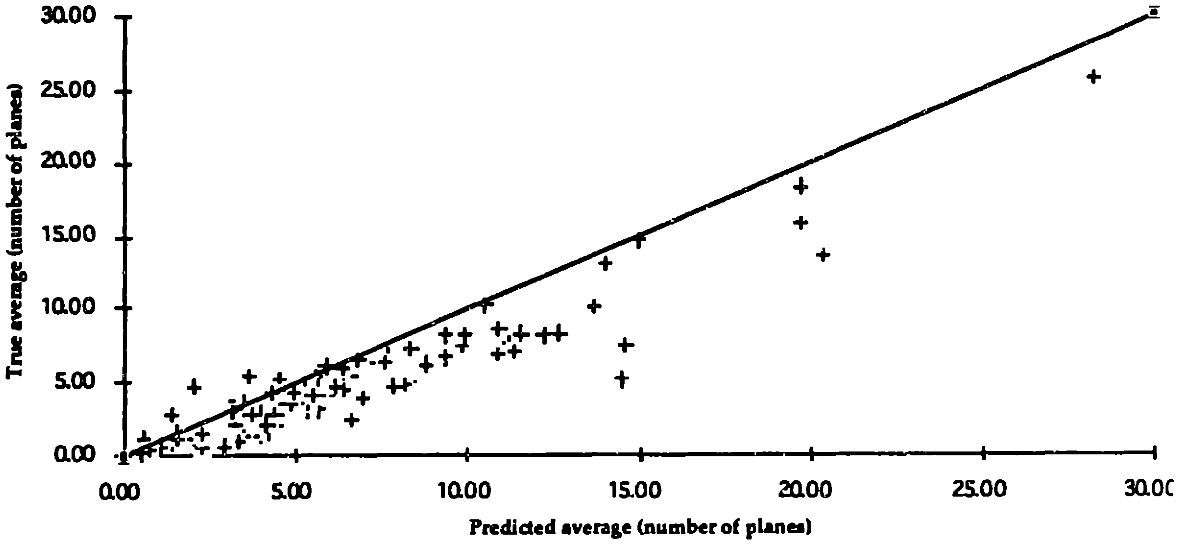
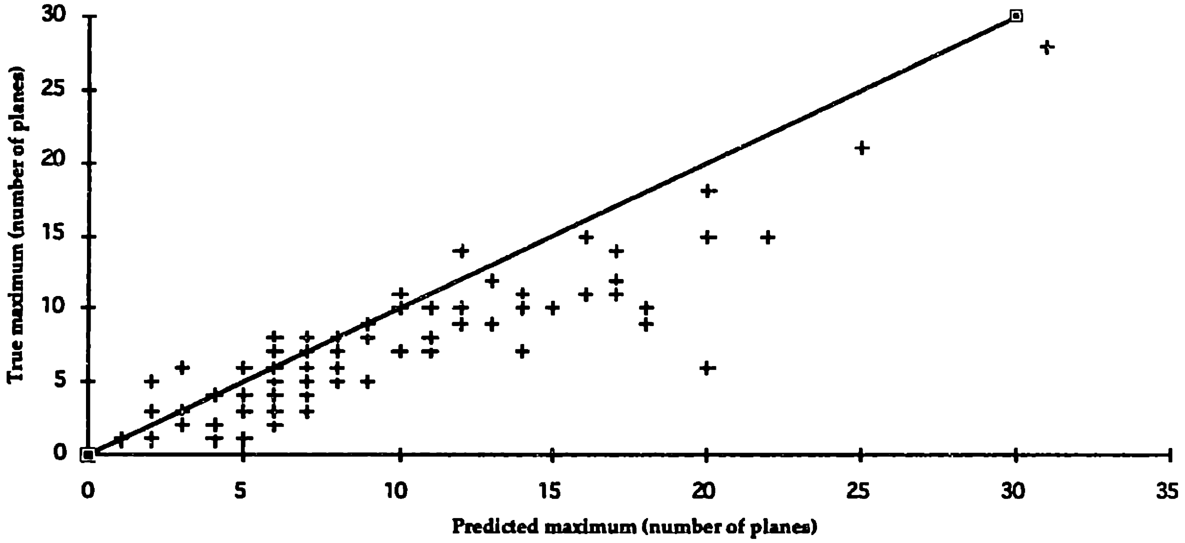


Figure 8

Plot of True 15-Minute Maximum versus Predicted 15-Minute Maximum for 96 Data Points from 12 Sectors



If every prediction were accurate, then all of the 96 points would lie on a 45-degree line, with an intercept of 0 and slope of 1. But as Figures 6, 7 and 8 show, many points were not on the diagonal line representing perfectly correct predictions. We performed regressions of actual on predicted data to gain a sense of how far off each criterion chart was from this ideal situation. For the maximum criterion the regression gave us a sense of the accuracy of the current calling of alerts. Moreover, comparisons among the three regressions gave us an additional sense of how vulnerable each of the three criteria are to existing levels of predictive inaccuracy. If, for instance, the average squared criteria was attractive theoretically, but it was especially prone to large predictive errors, then it would become less attractive. The regressions allowed us to evaluate how sensitive the criteria are to current levels of predictive inaccuracy.

However, regressions are only one way to measure predictive inaccuracy under each criterion. By minimizing the sum of squared vertical misses, the regressions tell us essentially the sum of squared error. To gain an alternate perspective, we calculated an average of percentage error, obtained from the following formula:

Average absolute percentage error

$$= \left(\frac{1}{96}\right)\sum_{i}^{96}\frac{|E_i - A_i|}{E_i}$$

E_i = expected sector demand
A_i = actual sector demand

This formula gives equal weight to the same percentage error throughout the data set, while regression tends to emphasize the highest predictions and/or the least accurate predictions. The two statistics together provide a panoramic view of how sensitive each criteria is to predictive error as well as illustrating how much of the plane-specific error translates into both false alerts and unnecessary alerts.

In that vein, we also calculated the percentage of false alerts and unnecessary alerts under the three criteria. Our data comes from many different sectors with many different alert thresholds; thus, as a first cut, we took the <u>top 10%</u> of the range of predictive values and assumed that those predictions represented situations when alerts were warranted. Predictions that were above the 10% cut-off but whose actual demands were below were deemed "false alerts", while predictions that were below the cut-off but whose actual demands were above were considered "surprise alerts". We then changed the cut-off to the

25th percentile of the predicted workload/stress levels to see if the results changed dramatically.

Our proportion of "false alerts" has one main drawback: it contains predicted alerts that controllers may have prevented because they took the warnings seriously and rerouted planes. Such outcomes reflect not failures of the system but the working of the system at its best. Thus our statistic cannot be unequivocally described as "system error". However, there is no such caveat associated with the proportion of "surprise alerts". Chapter 8, which follows, presents figures of the data, regression equations, and the results of the average percentage error calculations; it also discusses possible implications of each finding.

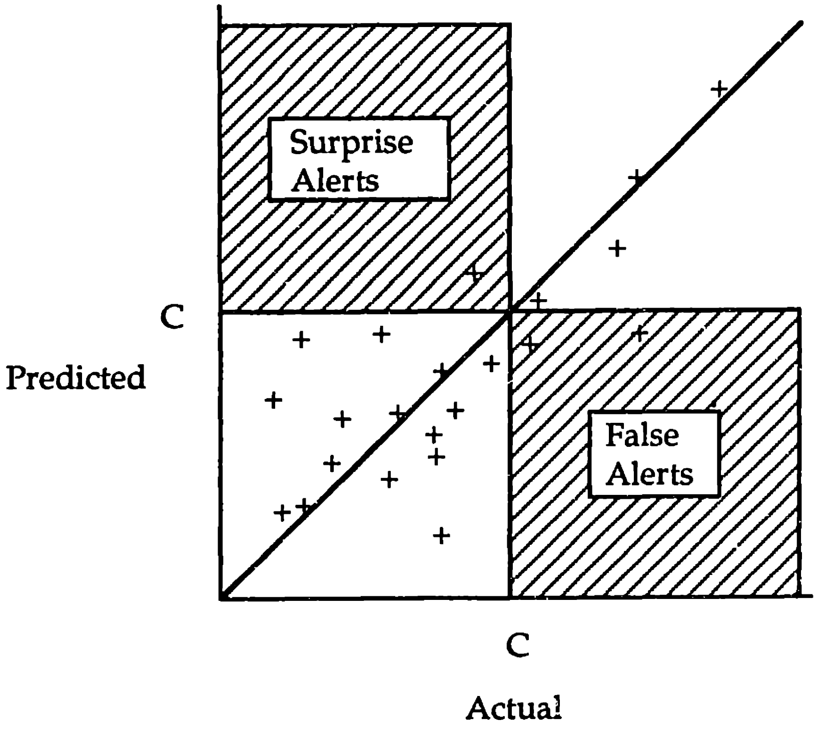
Chapter 8

The Accuracy of Alerts: Empirical Findings

The results in this chapter have two purposes: evaluating how ETMS plane-specific error translates into both "false alerts" and "surprise alerts", but also evaluating the maximum criterion by which alerts are called. Our first analysis resulted in Table 6 following, where the six numbers in the table represent the percentage of situations predicted to require alerts that do not actually materialize. For the 10% threshold level, these limits were determined by looking at the ninth-highest (ten percent of 96 observations is approximately 9) demand prediction and setting the limit so that data point and the nine larger data points qualified as "alerts". For example, if the 9th highest maximum demand prediction was 8, we would set the threshold at 7, so that all sectors containing a maximum demand of over 8 planes would be given alert status. Of course, these limits took different numerical values for each of the three criteria. And For the 25% threshold level we took the top 24 demand predictions for each criteria.

Because sectors had differing levels of traffic, this procedure weighted the experiences of busy sectors more than sectors which were less busy. In other words, our results are more pertinent for busier sectors, limiting their generalizability. Unfortunately, data limitations forced us to aggregate the twelve sectors, rather than perform a sector-by-sector analysis. We argue, however, that as a first cut this analysis gives a reasonable approximation of the proportion of "false" alerts and "surprise" alerts. (See Figure 9 for a graphical description of how "surprise alerts" and "false alarms" are defined.)

Figure 9: Graphical definitions of two types of alert inaccuracy for a given alert threshold C and hypothetical data.



At each of the two threshold levels, there were no "surprise" alerts called under any of the three criterion. This itself is surprising, because it means that out of our 96 observations, not one time did enough planes appear unanticipated into a sector to cause congestion that was unanticipated. The result concurs with our plane-specific result, where there were very few "surprise" planes compared to the many "no-shows".

Arguably, we should take comfort in this result, because surprise alerts are more dangerous than alerts that turn out to be false. By overestimating how many planes will be in a sector, the system may sometimes send controllers on "wild goose chases", but that seems less heinous than failing to identify potentially stressful situations. The only drawback to this conservatism is that the system may "cry wolf" enough that controllers fail to pay any attention to it, thereby rendering it useless.

Table 6. Proportion of alerts predicted that do not materialize, at two different threshold levels for three different alert criteria.

	Threshold Level	
	10%	25%
criterion		
maximum	0.73	0.52
average	0.33	0.44
root average square	0.56	0.50

There were, however, a significant number of "false alerts", as shown in Table 6. Two points can be garnered from this table. The first is that the proportion of false alerts, on average, across all of the different criteria, are surprisingly large. Close to 1 out of 2 alerts, on average, do not come to pass. Making sweeping generalizations about the accuracy of the system based on these numbers would be specious, given that we cannot separate system errors from alerts that were eliminated by controllers. For the maximum criterion, only

1 out of 4 alerts come to pass in the top ten percent of the demand distribution, while 1 out of 2 come to pass in the top 25 percent. Perhaps that outcome lends credence to the theory that controllers are working to lessen actual demand when the predicted is unacceptably high. However, we may reason that the drop-off in "false alerts" from the lower limit of the top quarter of the demand distribution range to the lower limit of the top tenth means that most of the false alerts in that range are due to system error, not to controller interference. We can estimate x, the proportion of false alerts in that range using the following formula:

$$\frac{0.73(9)+x(15)}{24}=.52$$

X is 0.38, meaning that our estimate of the number of alerts that were produced by system error is approximately 2 out of 5. that level of error is still very large. Table 6 also suggests that the average criterion is much less sensitive than the other two measures to system error in the upper tail of the demand distribution. That effect continues in the upper quarter of the distribution. The maximum criteria, by contrast, does not fare so well.

Regressions of Predicted Demand (P) on True Demand (T) for Three Criteria

Regressions with unconstrained intercept:

Maximum Criterion

$$T = 0.61 + 0.72P$$

$$\quad (1.48)\ (18.48)$$

$$R^2 = 0.784$$

Average Criterion

$$T = -0.18 + 0.79P$$

$$\quad (-0.73)\ (25.14)$$

$$R^2 = 0.871$$

Root Average Square Criterion

$T = -0.05 + 0.78P$

$(-0.18) \quad (24.02)$

$R^2 = 0.86$

Regressions with intercept not to fall below zero:

Maximum Criterion

$T = 0.61 + 0.72P$

$(1.48) \; (18.48)$

$R^2 = 0.784$

Average Criterion

$T = 0.79P$

(41.4)

$R^2 = 0.86$

Root Average Square Criterion

$T = 0.78P$

(41.5)

$R^2 = 0.86$

The average criterion also does well under the regression analysis, the results of which are above. Ideally, the slope of each regression would be 1, meaning that on average, the demand prediction equals the actual demand. We see that all three slopes are under one and the intercepts near or at zero, corroborating our theory that the system errs on the conservative side, preferring to call unnecessary alerts rather than miss any. The results also mesh with our previous plane-specific result that the probability of a "no-show" is higher than the probability of a "surprise" plane, while time-in-sector is predicted well by ETMS. Of the three slopes, the slope for the "average" criterion is closest to one.

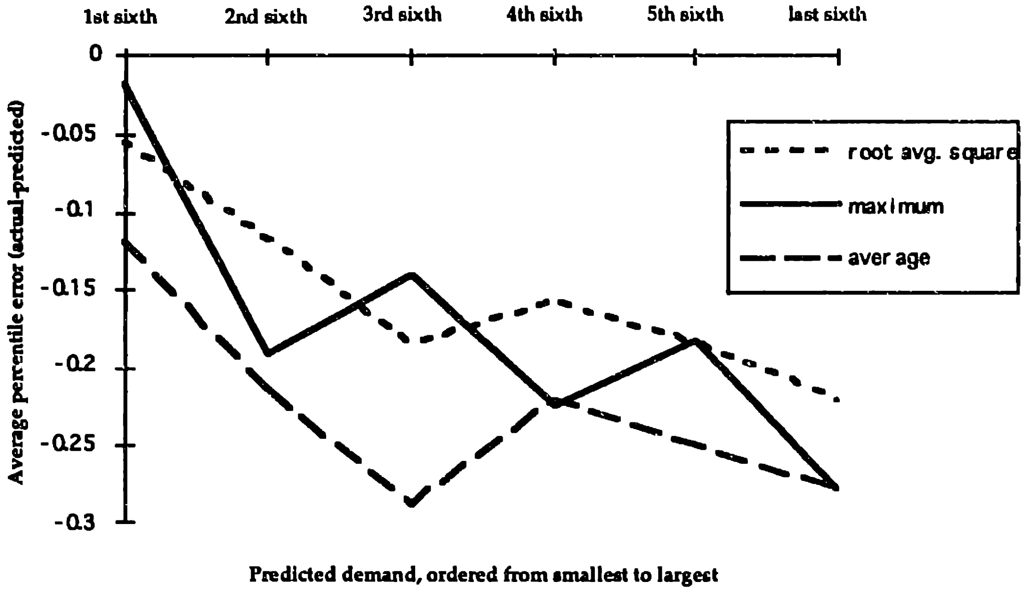
However, the most notable result of the first set of regressions is that the maximum criteria seems to have a poorer fit to the data than the other two criteria (R^2 is .77 for the ax, .87 for avg., and .86 for root avg. squared.) This result carries through to the second set of regressions as well (R^2 = .77, .86, .86 respectively). Yet the overall disparities across slopes are quite small.

As discussed in the previous chapter, regression tends to emphasize the highest predictions and/or the least accurate predictions. To avoid this emphasis, we calculated absolute percentage error. The absolute average percentage error statistics were 27%, 32%, and 22% for max., average, and root average squared criteria respectively. The root average square criteria has the least percentage error, where average had the best fit in the regression analysis. Either of these criteria is a clear winner, yet the maximum statistic is absent from the winner's table. It doesn't seem to be better than the other two in terms of accuracy.

Figure 10 illustrates how average percentage error behaves in each sixth of the demand distribution for each criteria. Again, we see that even when there are very few planes being predicted, and thus almost no chance that the planes will be re-routed, the system remains conservative, preferring to overpredict than underpredict the number of planes in a sector. The overall average percentage error statistics (not absolute percentage error) were -17%, -23%, and -15% for max., average, and root average squared criteria respectively. The average statistic suffers most from this overprediction, while the max. and root average statistics suffer comparatively less.

Figure 10:

Predicted Demand, Calculated Under Each of 3 Criteiia and Ordered From Smallest to Largest, Against Average Percentile Error



These results suggest that the system errs on the side of conservatism, and affirm that the limited errors in plane specific predictions translate into modest inaccuracies in aggregate plane "census" counts. They also provide reason to question whether the maximum criterion for calling alerts is the best rule – it does not stand out as the least sensitive to system error, nor is it obvious that it is the most theoretically sound of the three criteria.

Chapter 9

Conclusions

Here we present a summary of our results point-by-point, and then discuss some of their implications. We also identify what we did not do, and what questions further research might attempt to answer.

"Plane-Specific" Conclusions

1) The system has fewer "surprise" planes that spring up suddenly than "no-shows", or planes that are predicted to arrive at a sector but never show. (about 9% vs. 22% on average) The system also yields unbiased estimates of time-of-stay in sectors. Taken together, these outcomes mean that the system errs on the side of safety, tending to overestimate rather than underestimate the number of planes in any one place.

2) "No-show" prediction errors differ significantly across the three sector altitude levels: low, high, and super high. Midpoint and length-of-stay errors also differ across sectors. Predictions made for planes going through super-high sectors are excellent in terms of midpoint and length-of-stay error, yet relatively poor in that

there is a high percentage of "no-shows". "Surprise" prediction errors at first revealed no difference between sector altitudes, but when separated into commercial and non-commercial aircraft, differences did arise.

3) Prediction errors are substantially larger for aircraft still on the ground than for planes in the air; these predictions about planes not yet airborne are systematically too optimistic.

4) We also found that there is little correlation between either midpoint or length-of-stay error and predictive time-in-advance. There seems to be a minimum amount of error in predictions, but very little error is added by an increase in how far in advance a prediction is calculated. And this minimum error seems to be less than a minute for commercial planes in the air.

5) However, as the system looks further and further ahead, it tends to "miss" predicting the appearance of planes that actually enter a particular sector. For each ten minute increase in the time in advance the prediction is made, the system tends to miss an additional 2% of planes that actually entered. These additional planes become "surprise" visitors to the sector, unanticipated by the system.

6) For a given sector and plane, an ETMS prediction's absolute entry-time error is related to the absolute length-of-stay error. However, the information is difficult to use because if a plane enters a sector later than predicted, for instance, we anticipate an unusual length of stay but do not know whether it will be longer or shorter than normal.

7) Predictions for planes traversing a sector in a given direction during a particular period seem related to errors for planes going the same direction during the next period, but our efforts to capitalize on that relationship, have proven fruitless. It is possible that the relationship is too tenuous to realize any practical benefits.

Conclusions about Future Congestion

8) In our data set, there were no "surprise" alerts, meaning that no unacceptable congestion levels arose without warning. However, on average about half the alerts that were predicted to occur did not. We were unable to differentiate between congestion that was avoided by controller intervention and alerts that were unnecessary in the first place; however, by looking at alerts that were unlikely to have been alleviated by controllers, we estimate that perhaps 30% of those alerts are "false alarms".

9) The existing maximum criterion for calling alerts suffers greater deterioration in predictive accuracy because of ETMS predictive error than do criteria based on average or root average-squared rules. The maximum criterion did not emerge from our analysis as the clearly better or more effective than either of the other two criteria.

When the system errs, it errs on the side of safety. This is good news for the aircraft that rely on the system, because it seems that situations rarely arise in which many aircraft elude the ETMS's predictive eye and create havoc by unexpectedly arriving in as sector. On the other hand, by calling what seems to be a large number of unnecessary alerts, the system harms itself in two ways. If the ETMS "cries wolf", too many times, controllers get jaded and might rely less and less on the system. Also, the system can harm itself by creating the very stress it hopes to avoid. If controllers have to allocate appreciable time to investigating false alerts, the benefits of anticipating true congestion periods are eroded. The issue of false alerts may hurt the system more than is immediately obvious.

In that vein, two future projects might be able to provide information on how to minimize "false alerts" The first deals with "no-show" aircraft. We know that for aircraft in the air that do pass through the sector as predicted, the times of their visits are predicted with remarkable accuracy. The question is, what happens to the "no-shows," and how can the ETMS minimize their number? Even more important, perhaps, is addressing the fact that the predictions for aircraft on the ground tend to be consistently optimistic. Much of the error afflicting the overall system seems to stem from on-the-ground algorithms for modeling take-off times of planes; improving those algorithms might be the next major step in improving the system in general.

We have found that the system works well overall. There is room for improvement, as well as places where system designers should just leave the system alone. We tried to tie conclusions about how the system treats individual planes to conclusions about how accurately the system issues congestion alerts. We further suggested that the system seems safe for aircraft – but at perhaps the price of losing the confidence of the controllers. We see reducing the number of "false alerts" through a reducing the number of "no-show" aircraft as the best way to keep the system safe while earning the confidence of those who will use it.

Bibliography

MITRE Corporation, "A Preliminary Analysis of the Predicted Aircraft Position Accuracy of the Enhanced Traffic Management System Version 4.2", December 1992. McLean, VA.

Volpe National Transportation Center, "Enhanced Traffic Management System (ETMS) Functional Description Version 4.2", VNTSC-DTS56-TMS-003 , January 1992. U.S. Department of Transportation, Kendall Square, Cambridge MA.

Ming-Cheng Chiang, "A Study of Errors in Prediction Arrival Fix Times in Air Traffic Control", June 1993. Thesis submitted to the Department of Aeronautics and Astronautics in partial fulfillment of the Degree of Master of Science in Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.