2006

Coordination, Matchmaking, And Resource Allocation For Large-scale Distributed Systems

Xin Bai
University of Central Florida

ANALYSIS OF AIRCRAFT ARRIVAL DELAY AND AIRPORT ON-TIME PERFORMANCE

by

YUQIONG BAI
M.S. Tongji University, China, 2004
B.Tech. Huazhong University of Science and Technology, China, 2001

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil and Environmental Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Mohamed A. Abdel-Aty, Ph. D., P.E., Chair,
Xin Li, Ph. D.,
Chris Lee, Ph. D.

Spring Term
2006

ABSTRACT

In this research, statistical models of airport delay and single flight arrival delay were developed. The models use the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the Surface Airways Weather Data from the National Climatic Data Center (NCDC). Multivariate regression, ANOVA, neural networks and logistic regression were used to detect the pattern of airport delay, aircraft arrival delay and schedule performance. These models are then integrated in the form of a system for aircraft delay analysis and airport delay assessment. The assessment of an airport's schedule performance is discussed.

The results of the research show that the daily average arrival delay at Orlando International Airport (MCO) is highly related to the departure delay at other airports. The daily average arrival delay can also be used to evaluate the delay performance at MCO. The daily average arrival delay at MCO is found to show seasonal and weekly patterns, which is related to the schedule performance. The precipitation and wind speed are also found contributors to the arrival delay. The capacity of the airport is not found to be significant. This may indicate that the capacity constraint is not an important problem at MCO.

This research also investigated the delays at the flight level, including the flights with delay ≥0 minute and the flights with delay ≥15min, which provide the delay pattern of single arrival flights. The characteristics of single flight and their effect on flight delay are considered. The precipitation, flight distance, season, weekday, arrival time and the time spacing between two successive arriving flights are found to contribute to the arrival delay. We measure the time interval of two consecutive flights spacing and analyze its

effect on the flight delay and find that for a positively delayed flight, as the time space increases, the probability of the flights being delayed will decrease.

While it was possible to calculate the immediate impact of originating delays, it is not possible to calculate their impact on the cumulative delay. If a late departing aircraft has no empty space in its down line schedule, it will continue to be late. If that aircraft enters a connecting airport, it can pass its lateness on to another aircraft. In the research we also consider purifying only the arrival delay at MCO, excluding the flights with originating delay >0. The model makes it possible to identify the pattern of the aircraft arrival delay. The weather conditions are found to be the most significant factors that influence the arrival delay due to the destination airport.

ACKNOWLEDGMENTS

I would like to begin by thanking my advisor, Dr. Mohamed Abdel-Aty, whose unending patience, astute guidance and infinite insight and trust made this all a reality. Working under him was a wonderful learning experience.

To my committee, Dr. Xin Li and Dr. Chris Lee, thank you for taking time to read my thesis, for offering valuable suggestions and for serving on my committee. I would like to thank Dr. Xin Li and Dr. Chris Lee again for being part of the research group and helping me.

Thanks to Dr. Xiaogang Su and Dr. Xuedong Yan, for many useful suggestions and sharing his statistical genius with me. Dr. Anurag Pande and Martin also deserve thanks for their help as members of my research group. Thanks to Engy, Garda, Daniel, Ravi, and Hari, for helping me on everything, from code to coffee.

And of course, thanks to my husband, Xiaodong Zhang, for always being here. The support and understanding of family, friends and my colleagues gave me a tremendous spiritual boost that helped me achieve this goal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Research Motivation

With the great increase in air traffic comes a large increase in the demand for airport capacity. However, airspace and airport capacity cannot keep increasing at a rate necessary to match the rising demand. When an airport's capacity is reduced during "peak hours", the demand for an airport's resources exceeds the capacity that the airport can afford. This is known as a capacity-demand imbalance. Demand refers to the number of flights scheduled to arrive or depart in a given time period (rate of flight arrivals or departures). Capacity is the maximum number of flight arrivals or departures in a given time period. The direct result of the capacity-demand imbalance is the airport congestion and flight delay. Many major airports around the world have significant delay problems as a result of an imbalance between capacity and demand (Aisling and Kenneth, 1999).

Flight delays are obviously frustrating to air travelers and costly to airlines. Airline companies are the most important customers of the airport (Ashford and Wright, 1992). A well-known phase 'the airplane earns only when flying' holds true. On-time performance of airlines schedule is key factor in maintaining current customer satisfaction and attracting new ones. Flight schedule of the airport is the key to planning and executing airlines' operation (Wu, 2005). With each schedule, the airline defines its daily operations and commits its resources to satisfying its customers' air travel needs. Therefore, one of the basic requirements all airlines place on the ground handling is to ensure high efficiency of handling activities, avoiding delays (Mueller, et al., 2002).

Flight delay is complex to explain, because a flight can be out of schedule due to problems at the airport of origin, at the destination airport, or during the airborne. A combination of these factors often occurs. Delays can sometimes also be attributable to airlines. Some flights are affected by reactionary delays, due to late arrival of previous flight. These reactionary delays can be aggravated by the schedule operation. Flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. Even if complex, there exist some pattern of flight delay due to the schedule performance and airline itself. Some results extracted from the case study on Orlando International Airport (MCO) can help to better understand the phenomenon.

1.2 Problem Statement

Our case study is Orlando International Airport (MCO). The generality of a number of the findings may be limited, however, the methodology developed in this paper is widely applicable.

Orlando International Airport (MCO) is Florida's busiest airport, serving 56 airlines and around 30 million domestic passengers each year, with scheduled non-stop services to 84 US and 17 international destinations. More than 33 million passengers fly in and out of MCO each year, making it fourth busiest airport in the country for domestic travelers and the 14th in the country for total passengers (from http://www.orlandoairports.net).

The airport is presently moderately congested and for the past several years. While the domestic air traffic in MCO has greatly increased over the last 10 years, especially in 2004(14%) and in 2005(10%), it is predicted to continue to increase at a rate of 3 to 5% over the next 15 years, which has placed a heavier burden on air traffic control and airport facilities. Airport capacity will lose at the rate necessary to catch up with the rising demand. Because of the surge in air traffic and the limited capacity of airports, the capacity-demand imbalance will become more and more serious, which results in the airport congestion and flight delay.

What is more, the inherent randomness of air traffic systems cannot consider stochasticity enough in schedule planning. Because of this, there is often a notable discrepancy between a schedule and actual performance, which will increase the delay problem. It is vital that methodologies and tools be developed to analyze the increasing flight delay.

In air traffic flow management (ATFM), delay and congestion incur due to uncertainty of future landing capacity over a several hour interval. Ground holding program is one of the basic methods of lowering the cost of this problem. It means to have a flight wait on the ground at its point of origin than to have it circle the airport at its destination, unable to land.

If adverse weather conditions are anticipated at one airport the Federal Aviation Administration (FAA) issues a ground delay program (GDP) at this airport that increases the gap between successive flight arrivals to ensure safe operations. In most cases, the available slots for flight arrivals are less than what is required for the original planned schedule (Ball et al., 2000). Thus, a scheduled flight could be held at its origin, diverted to another airport or in the worst case it could be canceled. These disruptions in the planned flight schedule impact availability of crews and aircrafts for future flights. For instance, if a flight is delayed, its crewmembers may misconnect their next scheduled flights. They may also exceed the maximum allowed (legal) duty period length resulting in not completing remaining flights in their planned schedule (Yu et al., 2003).

Studies have identified the stages of flight in which delays occur and the causal factors that result in delays. For example, DOT classifies delays as gate delay, taxi-out delay, airborne delay and taxi-in delay. And the data shows that 84% of all delays occur on the ground (gate, taxi-out, taxi-in), out of which 76% are prior to takeoff (gate, taxi-out), suggesting that focusing on ground delay prediction will have the most impact on improving forecasting algorithms (Mueller, et al., 2002). So the arrival delay in this thesis is the delay value counting at the gate.

Empirical studies on airport congestion have identified several reasons which generate flight delays: saturation of airport capacity (including air transportation control activities), airline problems, reactionary delays, passengers and cargo, weather and other unpredictable disruptions (e.g. strikes). Among all these reasons, delay time experienced by flights and passengers can be mostly attributed to the first two groups: problems caused by air transportation control and airports, and by airlines. The impact of the most common and important of these factors will be discussed in chapter 3.

Inclement weather causes delays not only at airports experiencing the inclement weather, but also at airports with flights connecting from the airports experiencing inclement weather. During inclement weather, airport capacity is reduced due to increased aircraft separations. Because instrument landing systems are required for aircraft navigation in these conditions, this situation is called Instrument Meteorological Conditions, or IMC. Clear weather is known as Visual Meteorological Conditions or VMC.

In order to represent in our model this complex formation of flight delays, we will concentrate on three main reasons: airports' capacity, characteristics of individual flights, and weather conditions.

1.3 Research Objectives

On-time performance of airlines schedule is a key factor in maintaining current customer satisfaction and attracting new ones. However, flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, these delays could dramatically propagate over time and space unless the proper recovery actions are taken (Mueller, et al., 2002). This thesis presents models which projects individual arrival flight delays and alerts for possible future breaks during irregular operation conditions. Using the prediction model, it is possible to test sensitivity of overall schedule performance to the schedule time parameter.

Flight delay is a complex phenomenon. Even if complex, there exist some pattern of flight delay due to the schedule performance and airline itself. Due to the arrangement of airline schedule, the flight delay may show seasonal, weekly or daily patterns, and also show some preference according to airborne time, flight distance and origination areas etc. This is the interest of this thesis.

While it was possible to calculate the immediate impact of originating delays, it is not possible to calculate their impact on the cumulative delay. If a late departure aircraft has no empty space in its down line schedule, it will continue to be late. If that aircraft enters a connecting airport, it can pass its lateness on to other aircraft. In the research we also consider purifying only the arrival delay at MCO, excluding the flights with originating delay > 0. The model will make it possible to see the pattern of the aircraft arrival delay.

The analysis of an airport's schedule performance is another focus in this thesis. The airport delay distributions and the delay assessment of airport are presented. The results of our research show that the arrival delay is highly related to the departure delay at the originate airport. The patterns of daily average arrival delay at MCO are also carried out. Schedule design involves establishing a consistent rule for selecting the correct amount of time to allocate to each flight segment. In response to flight delay predictions and reason for these delays that are generated by the model, which can give indications for the appropriate recovery actions to recover/avoid these delays.

1.4 Organization of the Thesis

Following this introductory chapter, chapter 2 gives a background and description of flight delay along with a literature review of delay models, simulation methods and the statistical techniques used in the thesis.

Chapter 3 provides descriptions of the data sources and definitions of the data used to calibrate the statistical models of the thesis. There are two main data sources: the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the climatic data from the National Climatic Data Center (NCDC).

Chapter 4 presents the airport delay distribution and delay assessment. Then the average daily arrival delay models are carried out to analyze the airport arrival delay and pattern detection.

Chapter 5 presents models for delay analysis of individual arrival flights. Patterns between the flights with no delay and late flights are found. At the same time the patterns between the flights with low delay and high delay are found

In Chapter 6 we consider purifying only the arrival delay at MCO, excluding the flights with originating delay > 0. The model will make it possible to see the pattern of the aircraft arrival delay more clearly.

The final chapter consists of summary and conclusions from this research and provides insight into future research.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

2.1 Discussion of Flight Delay

Flight delay is a complex phenomenon, because it can be due to problems at the origin airport, at the destination airport, or during airborne. A combination of these factors often occurs. Delays can sometimes also be attributable to airlines. Some flights are affected by reactionary delays, due to late arrival of previous flights. These reactionary delays can be aggravated by the schedule operation. Flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. Even if complex, flight delays are nowadays measurable. And there exist some pattern of flight delay due to the schedule performance and airline itself (Wu, 2005). Some results extracted from the case study of Orlando International Airport (MCO) can help to better understand the phenomenon.

Two government agencies keep air traffic delay statistics in the United States. The Bureau of Transportation Statistics (BTS) compiles delay data for the benefit of passengers. They define a delayed flight when the aircraft fails to release its parking brake less than 15 minutes after the scheduled departure time. The FAA is more interested in delays indicating surface movement inefficiencies and will record a delay when an aircraft requires 15 minutes or longer over the standard taxi-out or taxi-in time (Mueller, et al., 2002).

Generally, flight delays are the responsibility of the airline. Each airline has a certain number of hourly arrivals and departures allotted per airport. If the airline is not able to get all of its scheduled flights in or out each hour, then representatives of the airline will determine which flights to delay and which flights to cancel (from http://www.travelforecast.com).

These delays take one of three forms, ground delay programs, ground stops, and general airport delays. When the arrival demand of an airport is greater than the determined capacity of the airport, then a ground delay program may be instituted. The airport capacity is unique to each airport, given the same weather conditions. The various facilities at an airport can determine how much traffic an airport can handle during any given weather event. Generally, ground delay programs are issued when inclement weather is expected to last for a significant period of time. These programs limit the number of aircraft that can land at an affected airport. Because demand is greater than the aircraft arrival capacity, flight delays will result.

Second, ground stops are issued when inclement weather is expected for a short period of time or the weather at the airport is unacceptable for landing. Ground stops mean that traffic destined to the affected airport is not allowed to leave for a certain period of time.

Lastly, there are general arrival and departure delays. This usually indicates that arrival traffic is doing airborne holding or departing traffic is experiencing longer than normal taxi times or holding at the gate. These could be due to a number of reasons, including thunderstorms in the area, a high departure demand, or a runway change. Our research finds that arrival and departure delays are highly correlated. Correlation between arrival and departure delays is extremely high (around 0.9 for 2002 and 2003). This finding is useful to prove that congestion at destination airport is to a great extent originated at the departure airport.

In order to understand flight delay, it is useful to consider the phenomenon of scheduled delay. The simplest way of reducing delays is not to increase the speed and efficiency of the system to meet the scheduled time, but to push back the scheduled time to absorb the system delays. As a result, one estimate put the number of scheduled delays that were built into airline schedules in 1999 at 22.5 million minutes. The number of arrival delays reported by BTS would have been nearly 25% higher in 1999 if airlines had maintained their 1988 schedules (Wu, 2005).

Sources of airport delay include many elements, such as weather, airport congestion, luggage loading, connecting passengers, etc. Weather is the main contributor to delays in the air traffic control (ATC) system. Traffic volume delays are caused by an arrival/departure demand that exceeds the normal airport arrival rate (AAR)/airport departure rate (ADR). The demand may also exceed the airport capacity if AAR and ADR are reduced due to weather conditions at the airport, equipment failure or runway closure. Delays may also be attributed to airline operations procedures (Aisling and Kenneth, 1999).

2.2 Literature Review

2.2.1 Literature on Delay Analysis and Potential Remedies

The increase in delays in the National Airspace System (NAS) has been the subject of studies in recent years. The literature on delay analysis and its potential remedies extends back over several decades. Levine (1969) argues that pricing is a better means of allocating scarce airport capacity to meet the demand than other mechanisms being considered at the time, such as slot allocation.

The Federal Aviation Administration (FAA) describes the increase in delays and cancellations from 1995 through 1999. Schaefer and Miller (2001) found that the current system for collecting causal data does not provide the appropriate data for developing strong conclusions for delay causes and recommend changes to the current data collection system.

Allan et al. (2001) examined delays at New York City Airports from September 1998 through August 2000 to determine the major causes of delay that occurred during the first year of an Integrated Terminal Weather System (ITWS) use and delays that occurred with ITWS in operation that were "avoidable" if enhanced weather detection. The methodology used in the study has considered major causes of delays (convective weather inside and well outside the terminal area, and high winds) that have generally been ignored in previous studies of capacity constrained airports such as Newark International Airport (EWR). The research found that the usual paradigm of assessing delays only in terms of Instrument Meteorological Conditions (IMC) and Visual Meteorological Conditions (VMC) and the associated airport capacities is far too simplistic as a tool for determining which air traffic management investments best reduces the "avoidable" delays.

Schaefer and Miller (2001) use the Detailed Policy Assessment Tool (DPAT) to model the propagation of delay throughout a system of airports and sectors. To estimate delays, throughputs, and air traffic congestion in a typical scenario of current operations in the U. S., DPAT models the flow of approximately 50,000 flights per day throughout the airports and airspace of the U. S. National Airspace System (NAS) and can simulate flights to analyze delays at airports around the world. They obtained results for local flight departure and arrival delays due to IMC, propagation for IMC, comparisons to

VMC results, and a comparison of propagated delays to entire system.

Rosen (2002) measures the change in flight times resulting from infrastructure-constant changes in passenger demand. Results indicate that delays rise with the ratio of demand to fixed airport infrastructure, decreasing average flight times by close to seven minutes after the sharp decrease in demand in the Fall of 2001. Flight time differences between the airlines in the sample are small, though the larger United had shorter average flight times in the winter quarter than America West, the smaller airline in the data sample.

Janic(2003) presents a model for assessment of the economic consequences of large-scale disruptions of an airline single hub-and-spoke network expressed by the costs of delayed and cancelled complexes of flights. The model uses the scheduled and affected service time of particular complexes to determine their delays caused by disruption.

During the last decade, a considerable attention has been given to proactive schedule recovery models as a possible approach to limit flight delays associated with Ground Delay Programs (GDP) (Abdelghany et al., 2004; Clarke, 1997). In these models, the impact of any reported flight delays, due to GDP or any other reason, is propagated in the network to determine any possible down-line disruptions (Monroe and Chu, 1995).

Wu (2005) explores the inherent delays of airline schedules resulting from limited buffer times and stochastic disruptions in airline operations. It is found that significant gaps exist between the real operating delays, the inherent delays (from simulation) and the zero-delay scenario. Results show that airline schedules must consider the stochasticity in daily operations. Schedules may become robust and reliable, only if buffer times are embedded and designed properly in airline schedules.

2.2.2 Review on Methodology of Delay Analysis

Suzuki (2000) proposes a new method of modeling the relationship between on-time performance and market share in the airline industry. The idea behind the method is that the passengers decision to remain (use same airline) or switch (use other airlines) at time t depends on whether they have experienced flight delays at time t-1 or not.

Air traffic flow management (TFM) (Ball, Connolly, and Wanke 2003) procedures such as Ground Delay program (GDP), Ground Stop (GS), or Miles-in-Trail (MIT) metering are options available to the Air Traffic Management (ATM) authority to manage airway congestion and to respond to anticipated weather conditions (Wanke et al. 2003). The effects of such complex interactions was quantified with either discrete event simulation or mathematical models or both. In this analysis, the authors developed a recursive MIT penalty function to quantify the ripple effects of specific MIT programs over relevant sets of flights and flight restrictions within the NAS. In conjunction with discrete event simulation, it is possible to examine and quantify the total impacts of various TFM programs for alternatives analysis and provide a comparison across several alternative TFM programs available to air traffic flow management decision-makers.

Hansen (2002) analyzes runway delay externalities at Los Angeles International Airport (LAX) using a deterministic queuing model. The model allows estimating the delay impact of each specific arriving flight on each other specific arriving flight. The research finds that, despite being only moderately congested (average queuing delay only 4 min per arriving flight), individual flights can generate as much as 3 aircraft-hours of external delay impact on other flights, with an average impact of 26 aircraft-minutes and

3400 seat-minutes. About 90 percent of this impact is external to the airline as well as the flight, a consequence of the lack of a dominant airline at LAX.

Shomik et al. (2002) presents an analysis of the possible impact of the application of slot controls as a demand management measure at San Francisco International Airport (SFO). A deterministic queuing model that uses an actual arrival schedule as input and simulates arrival delay based on available arrival capacity is used to estimate delay reduction potential of slot controls. The conclusions show the overall potential of slot controls to alleviate delay at SFO and their non-delay consequences.

Mehndiratta et al. (2002) propose a simulation framework to analyze the effects of stochastic flight delays on static gate assignments. The results of testing the framework on actual Chiang Kai-Shek airport (Taiwan) operations were good, showing that the framework could be useful for airport authorities to perform gate assignments.

Abdelghany et al. (2004) present a flight delay projection model, which projects flight delays and alerts for down-line operation breaks for large-scale airlines schedules. The results show that down-line schedule disruptions are proportional to the number of flights impacted by the GDP. Furthermore, in the recorded GDP instances, aircraft appears to be the reason for most flight delays predicted by the model.

Hansen and Hsiao (2005) analyze the recent increase in flight delay in the US domestic system by estimating an econometric model of average daily delay that incorporates the effects of arrival queuing, convective weather, terminal weather conditions, seasonal effects, and secular effects (such as half year). Results suggest that, controlling for these factors, delays decreased steadily from 2000 through early 2003, but that the trend reversed thereafter.

Hansen and Zhang (2005) investigated the interaction between LaGuardia Airport (LGA) and the rest of the aviation system by estimating simultaneous equations of average LGA and National Airspace System delay using two-stage least squares. The results demonstrate that the arrival delay impact of the Aviation Investment and Reform Act for the 21st Century (AIR-21) on LGA was in the form of increased Ground Delay Program (GDP) holding, and that while delay increased markedly under AIR-21 there were also observable improvements in the ability of LGA airport to handle traffic.

Hansen and Peterman (2004) use censored regression to analyze the delay impacts of the implementation of Traffic Management Advisor (TMA) metering at Los Angeles International Airport (LAX) in order to assess whether and how they have affected NAS performance. The results show that weather variables are not significant in the IMC models. In contrast to the IMC results, weather effects are significant under VMC. Temperature, visibility, and wind all have significant effects in at least one of the time periods. The presence of these weather effects under VMC suggests that, as a result of the greater flexibility of VMC separation rules, the performance of the airport is more responsive to changing conditions. It is notable that temperature is one the influential factors.

The current method of valuing delay in benefit-cost analysis is insufficient for determining the distributional impacts of a technology change on users because it fails to account for the shifts in where benefits occur and to which users. Kanafani et al (2004) propose a theoretical framework for evaluating the distributive effects of technology changes that requires a new approach to the evaluation of delay and understanding efficiency in light of the state of the system. The framework defines different categories of delay per flight and a method for calculating the cost of each type of delay by stakeholder recognizing that the airlines have different business strategies and therefore have different preferences. A case study based on a recent study of the benefits of the Integrated Terminal Weather Service (ITWS) demonstrates that a detailed investigation of the breakdown of delay into components can lead to more accurate delay cost accounting.

2.2.3 Conclusions of Review

Statistical models and simulation method are used to analyze flight delay, including deterministic queuing model, censored regression, and econometric model etc. But we can see that the analysis on delay are carried on macroscopical data or microcosmic data with only a few days. That is because of the huge data of flights every day. So here the flight delay are categorized into several level, and the logistic regression models are used here to better identify the delay pattern. In this thesis, studies on airport delay and delay influence on individual flight are carried out, using multiple regression model, logistic regression models, neural network models and tree model. The influencer related to aircraft, airline operations, change of procedures and traffic volume are also discussed. This paper will detect the pattern of delay from the airport level in which delays occur, give basic statistics on their magnitudes and frequencies.

The data used will be described in chapter 3, and the casual factors will be introduced too in chapter3. Chapter 4, 5, 6 will focus on the statistic models on airport delay and individual flight delay.

2.3 Statistical Background

2.3.1 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is used to study the effects of one or more independent (predictor) variables on the dependent (response) variable. Most commonly, ANOVA is used to test the equality of means by analyzing the total sum of squares (about the combined mean), which is partitioned into different components (due to model or due to random error). The formulas below depict the function of single-factor and multiple comparisons of ANOVA proposed by Girden (1992).

- Single-Factor or One-Way

A factor is an independent variable. Thus in single-factor ANOVA, the effects of only one independent variable are being tested. For single-factor ANOVA, each level of the factor is referred to as a treatment. The null hypothesis is equality of factor level means. The Single-Factor ANOVA model can be written as $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where Y_{ij} represents the i^{th} observation of the j^{th} factor level

$i = 1, \ldots n_j$, $j = 1, \ldots k$,

n_j is the number of observations for the j^{th} factor level, k is the total number of factor levels, μ is the overall mean of all factor level means, and α_j is called the effect of the j^{th} factor level.

The unknown parameters (μ, α_j) are usually estimated from the data using the method of ordinary least squares (OLS). In OLS, $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \mu - \alpha j)^2$ is minimized with respect to μ, α_1, α_2, ..., α_k. As stated earlier, the equality of factor level means are tested by analyzing the decomposition of overall variance (total sum of squares). The deviation

$Y_{ij} - \overline{Y}$, the difference between each observation and the overall mean can be decomposed into two components: deviation between each factor level mean and the overall mean; and the deviation of each observation around its respective factor level mean.

The ‘total sum of squares’ equals the sum of the ‘sum of squares due to model’ plus the ‘sum of squares due to random error’. Each sum of squares term divided by its associated degrees of freedom results in its mean square (MS). The F-value that is used as a test statistic is the ratio of the mean square of the model and the mean square error. Mean square of the model can also be thought of as the mean squared deviation between groups (treatments) and the mean squared error as the mean squared deviation within groups. Large values of the F-statistic lead to the rejection of the null hypothesis of the factor level means being equal.

- Multiple Comparisons

When the F-test rejects the null hypothesis that there exists an equality of means, the procedure of multiple comparisons allows one to determine where the differences lie while controlling the simultaneous confidence coefficient (1-α). In general, the procedure of multiple comparisons is used to determine if there exists statistically significant differences between two or more factor level means. Each comparison is known as a contrast, L, and is defined as $L = \sum t_j \mu_j$, where t_j satisfies the restriction $\sum t_j = 0$.

There are three common methods of multiple comparisons that are used: the Tukey Method, the Scheffe’ Method and the Bonferroni Method. The Tukey method should be used when the factor level sample sizes are equal and the multiple comparisons of interest are all pairwise comparisons. Scheffe’s method is the most general method in that it can be used regardless of whether or not the factor level sample sizes are equal and when all possible comparisons are sought. The Bonferroni method can be used irregardless of the factor level sample sizes, but only for a prespecified set of contrasts. The method that yields the greatest amount of precision of the confidence intervals depends on the type and amount of multiple comparisons being made.

2.3.2 Logistic regression modeling

Logistic regression belongs to the group of regression methods for describing the relationship between explanatory variables and a discrete response variable. A logistic regression is proper to use when the dependent is categorized and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects (Allison, 1999). Binary logistic regression is used when the dependent variable Y can only take on two values (such as low delay vs high delay). The equation 1 and 2 below depict the logit of binary logistic and multiple logistic regression model proposed by Allison (1999).

The probability that flight with high delay will occur or not is modeled as logistic distribution in Equation 1.

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{1}$$

The Logit of the multiple logistic regression model is given by Equation 2.

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n \tag{2}$$

where, $\pi(x)$ is conditional probability of a highly delayed flight, which is equal to the number of highly delayed flights divided by the total number of flights. x_n is independent variables. The independent variables can be either categorical or continuous, or a mixture of both. Both main effects and interactions can generally be accommodated. β_n is model coefficient, which directly determines odds ratio involved. The odds of an event are defined as the probability of the outcome event occurring divided by the probability of the event not occurring. The odds ratio that is equal to exp(β_n) tells the relative amount by which the odds of the outcome increase (O.R. greater than 1.0) or decrease (O.R. less than 1.0) when the value of the predictor value is increased by 1.0 units.

Maximum likelihood (ML) is the method for estimating the logit model for grouped data and the only method in general use for individual-level data. Maximum likelihood is a very general approach to estimation that is widely used for all sorts of statistical models. There are two reasons for this popularity. First, ML estimators are known to have good properties in large samples. And the sampling distribution of the estimates will be approximately normal in large samples, which means that you can use the normal and chi-square distributions to compute confidence intervals and p-values. The other reason for ML's popularity is that it is often straightforward to derive ML estimators when there are no other obvious possibilities. One case that ML handles very nicely is the data with categorical dependent variables (Allison, 1999).

2.3.3 Tree Classification Method

Decision trees build classification models based on recursive partitioning of data. Typically, a decision tree algorithm begins with the entire set of data, splits the data into two or more subsets based on the values of one or more attributes, and then repeatedly splits each subset into finer subsets until the size of each subset reaches an appropriate level. The entire modeling process can be represented in a tree structure, and the model generated can be summarized as a set of “if–then” rules. Decision trees are easy to interpret, computationally inexpensive, and capable of coping with noisy data. Therefore, the techniques have been widely used in various applications(The introduction to SAS Enterprise Miner Software, available from SAS).

In tree-structured representations, a node represents a set of data, and the entire data set is represented as a root node. When a split is made, several child nodes, which correspond to partitioned data subsets, are formed. If a node is not to be split any further, it is called a leaf; otherwise, it is an internal node. In this report, we deal with binary trees, where each split produces exactly two child nodes(32).

When a data point falls in a partitioned region, a decision tree classifies it as belonging to the most frequent class in that region. The error rate is the total number of misclassified points divided by the total number of data points; and the accuracy rate is one minus the error rate. The splitting attributes and their values in decision trees are determined by a sort-and-search procedure, in conjunction with an impurity measure(The introduction to SAS Enterprise Miner Software, available from SAS).

2.3.4 Neural networks

Artificial neural networks are alternative computation techniques that can be applied to solve category analysis problems. In this section we describe multi-layer perceptron (MLP) and Radial basis function (RBF) neural network that are most commonly used neural network architectures.

2.3.4.1 MLP neural network architecture

The MLP network is one of the most popular neural network architectures that fit a wide range of applications such as forecasting, process modeling, and pattern discrimination and classification. MLPs are feed-forward neural networks trained with the standard back-propagation algorithm. They are supervised networks so they require a desired response to be trained.

An MLP neural network has input layer, hidden layer and output layer along with input and output bias. The net input to hidden layer neurons is determined through inner product between the vector of connection weights and the inputs. The activation function is applied to this net input of hidden neurons and the weights from hidden to output layer are then used to get the output of the network. These weights are the parameter estimated during the supervised training process and are then used to 'score' unseen observations. The activation function of hidden neurons is non-linear in nature and is critical in the functioning of the neural network for it allows the network to 'learn' any underlying relationship of interest between inputs and outputs. An MLP neural network shown in Figure 2.1 from Christodoulou and Georgiopoulos (2001) has input layer of size K, a hidden layer of size J and output layer of size I along with input and output bias. In the MLP architecture the connections are of feed-forward type; it means that the only connections allowed between nodes are from a layer of a certain index to layers of higher index. (Neural Networks, http://www.statsoft.com/textbook/stathome.html)

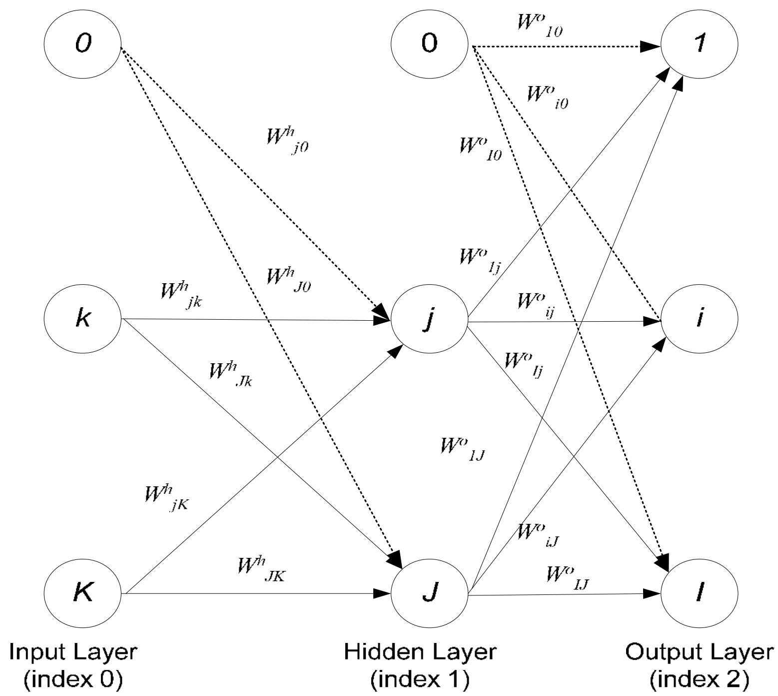

Figure 2.1 MLP neural network architecture

The number of input and output units is defined by the problem (there may be some uncertainty about precisely which inputs to use, a point to which we will return later. However, for the moment we will assume that the input variables are intuitively selected and are all meaningful). Once the number of layers, and number of units in each layer, has been selected, the network's weights and thresholds must be set so as to minimize the prediction error made by the network. This is the role of the training algorithms. (Neural Networks, http://www.statsoft.com/textbook/stathome.html)

The historical cases that you have gathered are used to automatically adjust the weights and thresholds in order to minimize this error. This process is equivalent to fitting the model represented by the network to the training data available. The error of a particular configuration of the network can be determined by running all the training cases through the network, comparing the actual output generated with the desired or target outputs. The differences are combined together by an error function to give the network error. The most common error functions are the sum squared error (used for regression problems), where the individual errors of output units on each case are squared and summed together, and the cross entropy functions (used for maximum likelihood classification).

In back propagation, the gradient vector of the error surface is calculated. This vector points along the line of steepest descent from the current point, so we know that if we move along it a "short" distance, we will decrease the error. A sequence of such moves (slowing as we near the bottom) will eventually find a minimum of some sort. The difficult part is to decide how large the steps should be.

Large steps may converge more quickly, but may also overstep the solution or (if the error surface is very eccentric) go off in the wrong direction. A classic example of this in neural network training is where the algorithm progresses very slowly along a steep, narrow, valley, bouncing from one side across to the other. In contrast, very small steps may go in the correct direction, but they also require a large number of iterations. In practice, the step size is proportional to the slope (so that the algorithms settles down in a minimum) and to a special constant: the learning rate. The correct setting for the learning rate is application-dependent, and is typically chosen by experiment; it may also be time-varying, getting smaller as the algorithm progresses (Christodoulou and Georgiopoulos, 2001) .

The algorithm therefore progresses iteratively, through a number of epochs. On each epoch, the training cases are each submitted in turn to the network, and target and actual outputs compared and the error calculated. This error, together with the error surface gradient, is used to adjust the weights, and then the process repeats. The initial network configuration is random, and training stops when a given number of epochs elapses, or when the error reaches an acceptable level, or when the error stops improving.

The details of the algorithm would make this point clearer and are provided below from Christodoulou and Georgiopoulos (2001). The objective function takes the following form:

$$F(w) = \sum_{p=1}^{P} \left[\sum_{k=1}^{K} (d_{kp} - o_{kp})^2 \right] \qquad (3)$$

where $\boldsymbol{w} = [w_1 \quad w_2 \quad \dots \quad w_N]^T$ consists of the interconnection weights in the network, d_{kp} and o_{kp} are the desired and actual values, respectively, for k_{th} output and p_{th} pattern. N is the total number of weights, P is the number of patterns, and K is the number of network outputs. The above equation may be rewritten as

$$F(w) = E^T E \qquad (4)$$

$$E = [e_{11} \cdots e_{K1} e_{21} \cdots e_{K2} \cdots e_{1P} \cdots e_{kP}]^T, \quad e_{kp} = d_{kp} - O_{kp} \quad k = 1, \cdots K, \quad p = 1, \dots, P$$

where E is the cumulative error vector (for all patterns).

2.3.4.2 Radial basis function (RBF) neural network

The RBF network is a popular alternative to the MLP, which although it is not as well suited to larger applications, can offer advantages over the MLP in some applications. For example, an RBF network can be easier to train than an MLP network. RBF networks have a number of advantages over MLPs. First, as previously stated, they can model any nonlinear function using a single hidden layer, which removes some design-decisions about numbers of layers. Second, the simple linear transformation in the output layer can be optimized fully using traditional linear modeling techniques, which are fast and do not suffer from problems such as local minima which plague MLP training techniques. RBF networks can therefore be trained extremely quickly (Neural Networks, http://www.statsoft.com/textbook/stathome.html).

Radial-Basis Function Networks contains an input layer, a hidden layer with nonlinear activation functions and an output layer with linear activation functions. A radial basis function network (RBF) has a hidden layer of radial units, each actually modeling a Gaussian response surface. Since these functions are nonlinear, it is not actually necessary to have more than one hidden layer to model any shape of function: sufficient radial units will always be enough to model any function.

In the hidden layer of an RBF, each hidden unit takes as its input all the outputs of the input layer x_i (Christodoulou and Georgiopoulos 2001). The hidden unit contains a "basis function" which has the parameters "centre" and "width". The centre of the basis function is a vector of numbers of the same size as the inputs to the unit and there is normally a different centre for each unit in the neural network. The first computation performed by the unit is to compute the "radial distance", d, between the input vector x_i and the centre of the basis function, typically using Euclidean distance:

$$d = \sqrt{(X_1 - c_1)^2 + (X_2 - c_2)^2 + \ldots + (X_n - c_n)^2}$$

The unit output a is then computed by applying the basis function B to this distance divided by the width w: $a = B(d / w)$

In feed forward neural network architectures the activation function of hidden neurons is applied to a net single value that is obtained by combining input vectors with the vector of connection weights between input layer to hidden layer. The function that combines the inputs with the weights may be referred to as the 'combination function'. In the MLP neural network architecture the combination function was simply the inner product of the inputs and weights.

There are two distinct types of Gaussian RBF architectures. The first type, the ordinary RBF (ORBF) network, uses the exponential activation function, so the activation of the unit is a Gaussian "bump" as a function of the inputs. The second type, the normalized RBF (NRBF) network, uses the softmax activation function, so the activations of all the hidden units are normalized to sum to one.

Note that the output bias has no role in an NRBF network since the constant bias term would be linearly dependent on the constant sum of the hidden units due to the softmax activation. The distinction and advantages of NRBF networks over the ORBFs are discussed in detail by Tao (1993). It was argued by Tao (1993) that the normalization not only is a desirable option but in fact is imperative.

In NRBF networks one may add another term to the Gaussian combination function referred to as the 'altitude' which determines the maximum height of the Gaussian curve over the horizontal axis. Based on the two parameters (width and height) defining the shape of combination function the NRBF networks may be categorized into five different types:

NRBFUN: Normalized RBF network with unequal widths and heights

NRBFEV: Normalized RBF network with equal volumes (a_i=w_i)

NRBFEH: Normalized RBF network with equal heights (and unequal widths) (a_i=a_j)

NRBFEW: Normalized RBF network with equal widths (and unequal heights) (w_i=w_j)

NRBFEQ: Normalized RBF network with equal widths and heights (a_i=a_j) and (w_i=w_j) where w_i and a_i represent the widths and heights, respectively, of the neurons in the hidden layer. Note that the last four categories of networks are special cases of the first and are more parsimonious in nature. It essentially means that with certain assumptions about the shape of the combination functions they reduce the number of parameters to be estimated.

The output activation function in RBF networks is customarily the identity. Using an identity output activation function is a computational convenience in training, but it is possible and often desirable to use other output activation functions just as you would in an MLP. The Neural Network node sets the default output activation function for RBF networks the same way it does for MLPs.

As mentioned earlier, training of RBFs takes place in distinct stages. First, the centers and deviations of the radial units must be set; then the linear output layer is optimized. Centers should be assigned to reflect the natural clustering of the data. The two most common methods are:

Sub-sampling. Randomly-chosen training points are copied to the radial units. Since they are randomly selected, they will represent the distribution of the training data in a statistical sense. However, if the number of radial units is not large, the radial units may actually be a poor representation (Haykin, 1994).

K-Means algorithm. This algorithm (Bishop, 1995) tries to select an optimal set of points that are placed at the centroids of clusters of training data. Given K radial units, it adjusts the positions of the centers so that each training point belongs to a cluster center, and is nearer to this center than to any other center and each cluster center is the centroid of the training points that belong to it.

Once centers are assigned, deviations are set. The size of the deviation (also known as a smoothing factor) determines how spiky the Gaussian functions are. If the Gaussians are too spiky, the network will not interpolate between known points, and the network loses the ability to generalize. If the Gaussians are very broad, the network loses fine detail. This is actually another manifestation of the over/under-fitting dilemma. Deviations should typically be chosen so that Gaussians overlap with a few nearby centers.

CHAPTER 3 DATA DESCRIPTION AND RELATED ISSUES

The analysis in this thesis is based on data from the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the climatic data from the National Climatic Data Center (NCDC). In the following chapters, models will be presented for the estimation of a vector of airport daily arrival delay and single flight arrival delay. These models are formulated using flight delay parameters and weather conditions at Orlando International Airport (MCO).

The delay statistics with the data specific to MCO airport is the subject of this section. A comprehensive characterization and comparison of the arrival and departure delay distributions will be presented. Historical delay data for the airport are summarized. To enable such an analysis, several data fields for every aircraft arriving at MCO airport from 2002 to 2003 were extracted from the database. The various causal factors related to aircraft, airline operations, weather and traffic volume are also discussed in section 3.3.

3.1 Airline On-Time Performance Data and Surface Airways Weather Data

The statistical models in this thesis are estimated on this data consisting of domestic flights with the destination of MCO. The data for the non-stop flights on scheduled service by certificated carriers to MCO were obtained from the Airline On-Time Performance Data from the Federal Aviation Administration (FAA). The data is collected by the Office of Airline Information, Bureau of Transportation Statistics (BTS).

The models are presented for the estimation of a vector of airport average daily arrival delay or single flight arrival delay. The data used consists of the non-stop domestic flights on scheduled service by certificated carriers with the destination of MCO. This database contains departure delays and arrival delays for non-stop domestic flights by major air carriers, and provides such additional items as origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi out and taxi in times, airborne time, and non-stop distance. The flight data used is from 01/01/2002 through 12/31/2003 excluding the cancelled and diverted flights. The following data fields were extracted for each aircraft in the database:

- identification code,
- information of the date of departure,
- original airport code,
- destination airport code,
- scheduled time and actual time of departure,
- scheduled time and actual time of arrival,
- scheduled flight time and actual flight time,
- arrival delay and departure delay,
- flight distance,
- cancelled and diverted flights.

Arrival performance and departure performance in this thesis is based on arrival at the gate. So the delay considered is based on the delay at the gate. The flights that leave the gate more than fifteen minutes after the scheduled time shown in the carriers' computerized reservations systems (CRS) are considered “late” while all other flights are recorded as “on-time”. Similar to origin airports, flights at destination airports are defined as “on-time” if they arrived at the gate within fifteen minutes of the scheduled time shown in the carriers’ CRS, while all remaining flights are defined as “late”. The delayed flights considered here are the flights with delay equal to or more than 15 minute.

Considering the number of random events impacting on-time arrival performance of the airline, it is surprising that the airlines can run on schedule at all. Recent statistics show that the airlines in MCO, in Table 3.1, arrive on time only between 60 and 70 percent of the time from 2000 to 2004, while there is no big increase in the air traffic volume. According to Table 3.1, there is an increasing trend in on-time performance from 2000 to 2003 excluding 2004, due to the hurricane impact in 2004 on MCO. For the analysis on the single flight delay, our interest is in the positively delayed flights and the on-time performance of the schedule, the before- time flights will not be considered, and only the on-time flights and positively delayed flights are considered

Table 3.1 On-time Performance at MCO

year	before_time		on_time		delay		Total
2000	11370	11.37%	64713	64.73%	23892	23.90%	99975
2001	14671	14.87%	65470	66.37%	18499	18.75%	98640
2002	12527	15.37%	54812	67.27%	14141	17.36%	81480
2003	13153	14.24%	63813	69.09%	15391	16.66%	92357
2004	12801	11.87%	73574	68.20%	21496	19.93%	107871

Weather conditions play a very important role in determining airport capacities. Since it is our goal to predict and estimate arrival flight delay, it is imperative that we consider weather conditions. Most often, weather-related flight delays are due to the interaction of two factors. One, how many planes can an airport accept during a given time period based on the weather (airport capacity)? Two, how many planes are scheduled to arrive (airport demand) during the same given time period?

There are a whole slew of weather parameters than cause flight delays. The most significant and common weather variables that cause delays are low clouds and low visibility. Low visibility may be due to fog, haze, smoke, and falling precipitation. When these conditions occur, planes may be spaced farther apart, thus resulting in fewer planes landing in any given hour. Wind, another typical factor, can also have a significant impact on flights. Strong low-level winds or wind shear may require that planes be spaced farther apart. Strong crosswinds may make some runways unusable.

We control for adverse weather using information about the amount of rainfall for every day at MCO. That database contains daily observations about the inches of the rainfall indicating whether the station had heavy rainfalls during that day. We also include in our model the daily average wind speed at Orlando International Airport. Airport wind speed and precipitation at MCO are the two weather variables from Surface Airways Hourly Weather Data, which is collected and archived by the NCDC.

3.2 Variables information

Since the interest of this research is on the arrival flight delay, we will consider the average daily arrival delay at MCO and the single arrival flight delay. The independent and dependent variables are introduced as below.

(1) Flight arrival delay

Arrival delay equals the difference of the actual arrival time minus the scheduled arrival time. The single flight delay metric uses directly the arrival delay for each flight with target delay level (delay ≥0 or delay≥15 min). While the average daily delay metric reflects all the arriving flights, flights that arrive early are assigned zero delay in the calculation. For the delay metric, *d(t)*, we used the average daily positive delay for all scheduled and completed flights from other airports to MCO airport in Orlando. It is the average of positive delay per flight per day.

(2) Maximum hourly flow rate (airport capacity)

The airport capacity refers to the ability of the various facilities in the airport system in handling the aircrafts' activities in the airport. The critical factor of the capacity is the relationship between the demand and capacity and how the transportation system's service time is affected. As service time increases, system delay may increase and overall system reliability decreases.

The preferred measure of the airport capacity is the ultimate or saturation capacity, which gives the maximum number of aircrafts that, can be handled during a certain period under conditions of continuous demand [2]. It is usually expressed in terms of operations per hour (arrivals or departures). This hourly capacity is the maximum number of operation that can be handled in a one-hour period under specific operating conditions, in particular, weather conditions (ceiling and visibility), air traffic control, the aircraft mix and the nature of operations. In this research, the airport capacity is represented by the maximum hourly capacity in one day (including arrival and departure flights) according to the actual departure time.

(3) Arrival demand

The arrival demand is included as another variable that may capture the incidence of congestion in the airport system. The arrival demand vector is represented by the sum of completed arrival flights to MCO in the hour when the flight occurs according to the scheduled arrival time.

(4) Flight distance

The effect of flight distance is captured by the categories of the flight distance, which respectively represent the flight distance of 0 to 750 miles, 750 to 1000 miles and greater than 1000 miles. These classes are categorized with the same percentage of total flights. These factors in the model can show whether the long-term flights influence the arriving delay more or the short-term flights.

(5) Spacing of Inter-arrival time

The space is used here another variable, which means the intervals between two consecutive arriving flights. This inter-arrival time is calculated as the time between each flight and the before flight according to the schedule arriving time. The aim of spacing is to find the relationship of delay and the schedule operation of the airport. The intervals between arrival flights differ under different weather conditions, runway conditions and operating conditions. If the traffic flows arrive smoothly, we can say that the spacing is constant and small which means the waiting and service time for each flight is controlled at a low level.

(6) Airport precipitation

The database contains daily observations about the inches of the rainfall indicating whether the station had heavy rainfalls during that day. That indicated that the adverse weather contributes to the delay in that day. The precipitation is in hundredth of an inch of rainfall per hour.

(7) Airport wind speed

The arrival delay can be affected by windy conditions, either because of the direct effect of the wind or because of associated conditions such as wind shear. The variable of daily average wind speed at MCO from NCDC is used in our model. The wind speed is speed of wind in mph per day.

(8) Seasonal variables

Seasonal effects are captured by a set of dummy variables. The seasonal variables indicate the seasons when the flights are scheduled. The season has four classes, spring (March-May), summer (June-August), fall (September to November), and winter (December to February), which represent the four seasonal patterns in Orlando. The model includes 3 seasonal dummies. These variables capture the seasonal changes in the flight delay.

(9) Weekly variables

These variables indicate the weekday when the flights are scheduled to detect the weekly pattern of delay.

(10) Origin airport regional variables

There are 82 airports that have direct flights to Orlando. Here the areas of the origin airports are divided into four parts that are southeast, southwest, northeast and northwest (appendix A) to detect the region impact on the arrival delay.

(11) Time effect

The time effect is defined by a set of dummy variables. These variables indicate the scheduled arriving time of each delayed flights. Here we classify the scheduled arrival time into 3 classes: morning, afternoon and evening. They are 7am to 11:59 am, 12 am to 4:59 pm, and 5 pm to 11:59 pm. So the model includes 2 dummies to capture the scheduled arrival time of each delayed flight.

(12) Interaction effects

In addition to the various factors above, we also investigated the interaction effects between season and time, time effect and weekend effect, weekend effect and distance effect by adding a set of variables with the corresponding interaction.

3.3 Airport Arrival Delay Distributions (including early arrivals)

In this section, air traffic delay characteristics at MCO were examined, and the focus is on aggregate statistics derived from the complete dataset, which includes all the traffic over the two-year period. The cumulative distribution of arrival delay and departure delay will be examined, along with other arrival delay characteristics.

3.3.1 Cumulative Distribution of Arrival Delay

Figure 3.1 and Figure 3.2 show individually the percentage of aircraft as a function of departure and arrival delays for individual flights. The y axle shows the percentage of aircraft, out of all the aircraft that arrived at MCO from 2002 to 2003, which had the number of minutes of delay as x axle. The negative value shows that the flight arrived before scheduled time, and the positive value shows the flight arrived after scheduled time. The value of zero shows the flight arrived exactly on time, which presents a small part of the arriving aircrafts. Note that arrival delays are computed at MCO while the departure delays are computed elsewhere (destination airports).

Based on the mean and standard deviations derived from the raw data, for the departing flights, 94% of the delays ranged from −10 to 53 min, with the mean being 7.881 minutes and the standard error (STD), 22.51 min. For the arriving flights, 94% of the delays ranged from −29 to 58 min, with the mean being 4.847 minutes and the STD, 25.82 min. The mean of departure (7.881 minutes) delay shows higher than the mean of arrival delay (4.847 minutes). From the distribution of departure delay and arrival delay, we can also see that the departure delay has a much larger part of zero value and smaller part of negative value compared to arrival delay, which indicates a large amount of aircrafts departure on time and a small amount of aircrafts departure before time.

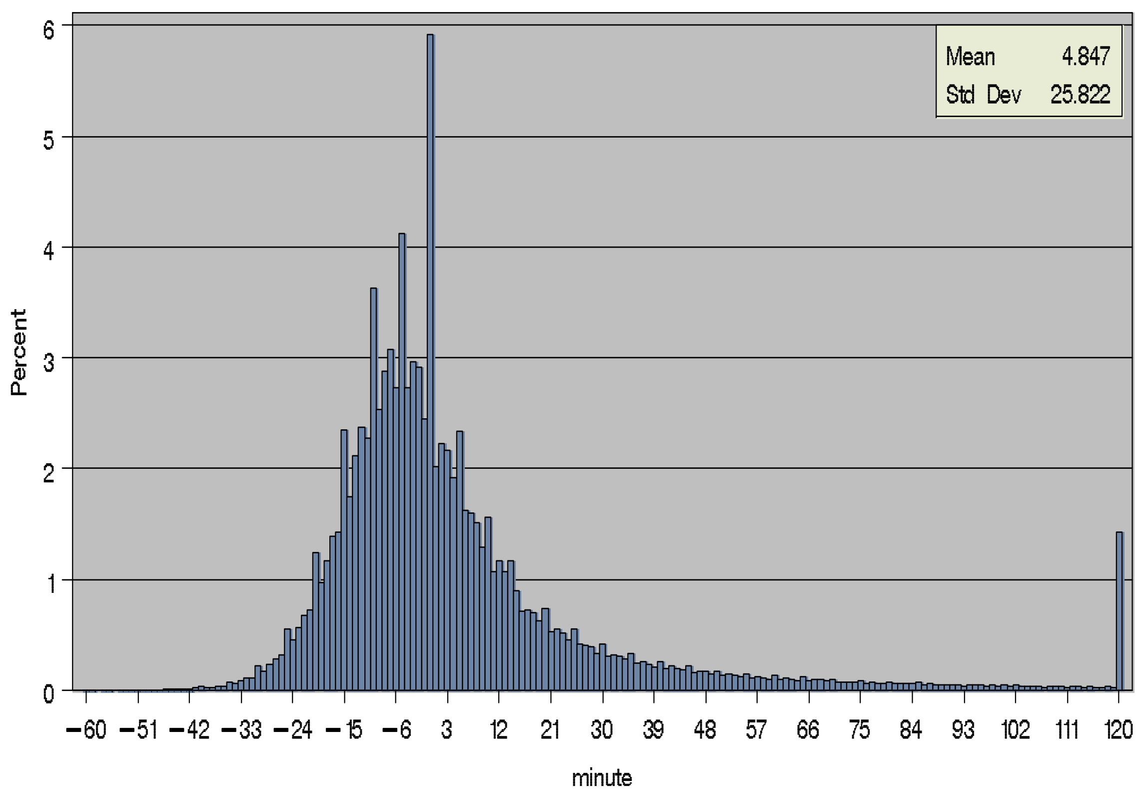

Figure 3.1 Percentages of aircraft as a function of arrival delays

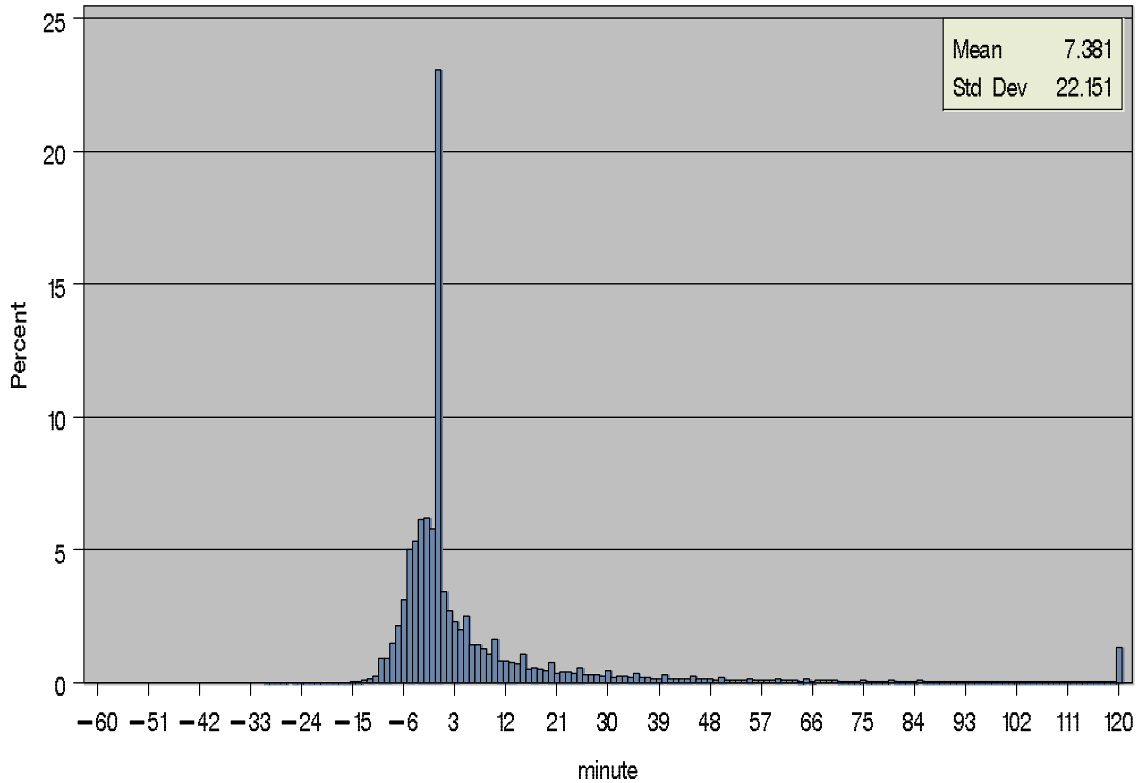

Figure 3.2 Percentages of aircraft as a function of departure delays

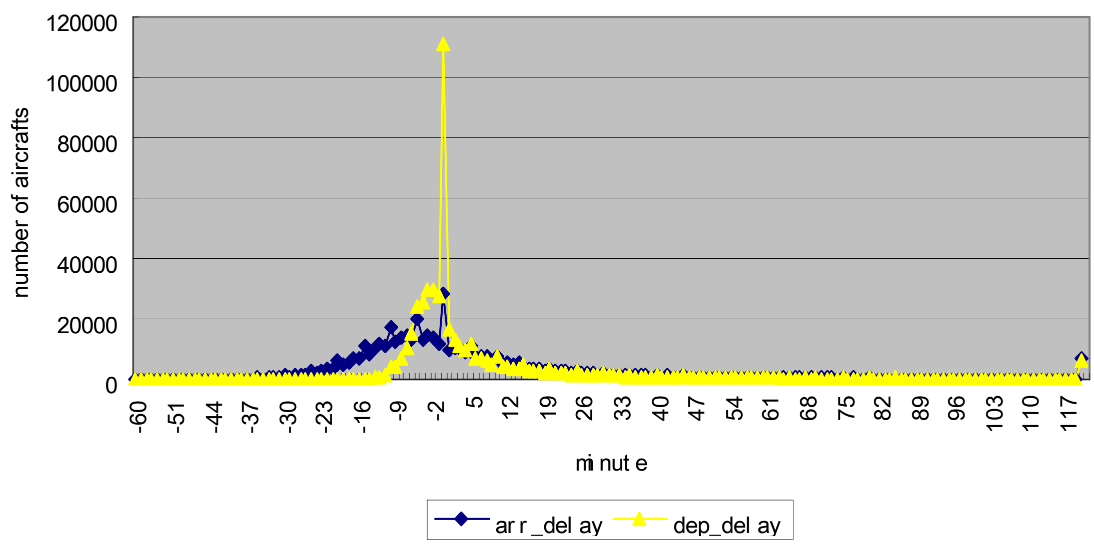

Figure 3.3 Number of aircrafts according to the distributions of departure delays and arrival delay

In Figure 3.3, the number of aircrafts as a function of the distributions of departure delays and arrival delay are showed. The light line shows the number of aircrafts as a function of departure delay minutes, and the dark line show that of arrival delay minutes. Observe from the two lines that a slightly greater percentage of aircraft encounter arrival delays than experience departure delays. This may be due to those aircrafts that experienced departure delays, which propagate through to become arrival delays, and those small number that did not experience departure delays but were subject to enroute delays or terminal delays, becoming arrival delays. It should be noted that the difference between the percentages of delayed departures and arrivals is rather small. And from Figure 3.1 and 3.2, it is evident that the percentages of delayed departures and arrivals are similar in some cases, suggesting that delay is frequently incurred on departure and carries through to arrival.

All these imply that most of the delay originates before departure, in another words, the departure delay has a high relation with arrival delay.

Considering the distribution of arrival delay and the definition of delay, we classify in this thesis 3 categories of delay. No delay means the delay time from 0 to 14 minutes, low delay represents the delay time from 15 to 29 minutes, and high delay is the delay time greater than or equal to 30 minutes. The notations "no delay", "low delay" and "high delay" here are relative and only for comparison purposes. In the real world, the airport authority should generate delay distributions suitable to its own operations for testing.

3.3.2 Arrival delay pattern over time

Airline travelers have become familiar in recent years with frequent delays and congestion. Air traffic has become concentrated particularly at certain seasons, at certain time of a week, or at certain times of the day.

Observe from Table 3.2 that in summer MCO bears the highest positive arrival delay, and a large part of delayed flights are with delay more than 15 minutes. In Figure 3.4, the seasonal pattern of delayed flights is explained more expressly. The air traffic has no big difference in four seasons (spring and summer bear heavier traffic volume), while in summer more flights are positively delayed and with delay of more than 15 minutes. And in fall, the delayed flight is much lower compared to other three seasons. In another word, the arrival delay at MCO shows seasonal pattern.

The delay concentration at certain season may depend on the area characteristics or the weather conditions. The adverse weather in summer at MCO may attribute to the arrival delays. And the travel pattern at MCO shows the summer and winter are the busiest seasons, especially for domestic travelers. The imbalance of capacity and demand may cause the arrival delay.

Table 3.2 Numbers of arrivals as a function of seasons

Season	Numbers of arrivals with delay≥15 min	Numbers of arrivals with positive delay	Number of arrivals
Spring	22348	50885	122925
Summer	29159	57100	123577
Fall	16789	41570	116343
Winter	25123	52881	117478

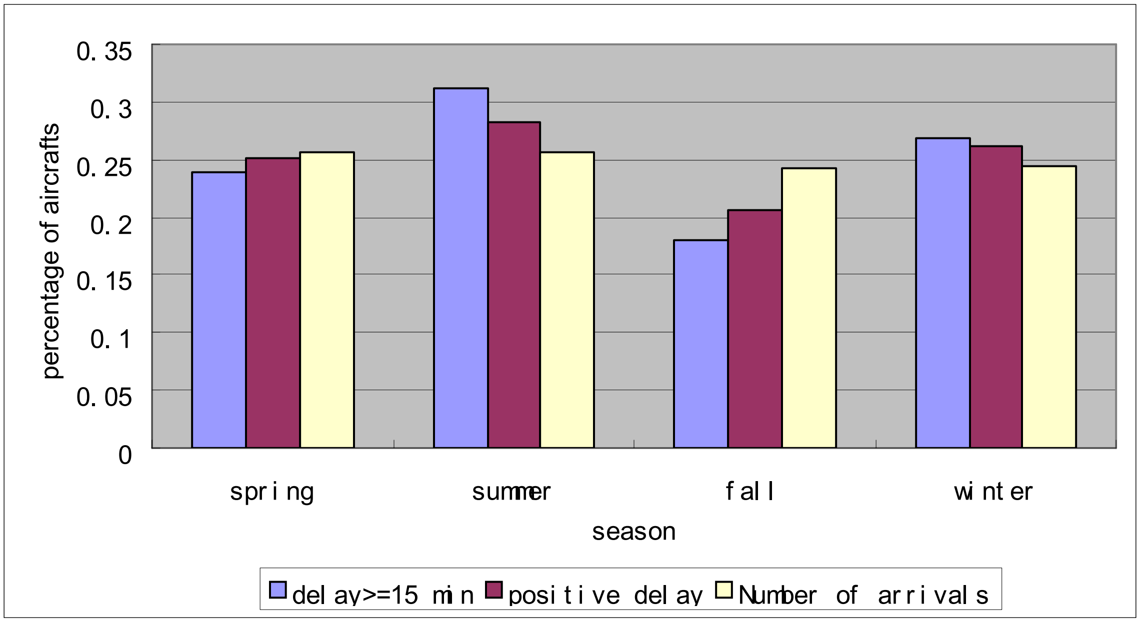

Figure 3.4 Delay distributions of arrivals as a function of seasons

In Figure 3.5, the blue bar represents the percent of aircrafts with positive delay, and the red one shows the percentage of traffic volume in one week. Although the weekends (Saturday and Sunday) have higher traffic volume, the variation from day to day is small. Observe from Figure 3.5 that although the traffic volume is lower on Thursday and Friday, the percentage of delayed aircraft is higher. A likely explanation for such a trend is the small variation in departures from day to day, which is related to the airline schedule.

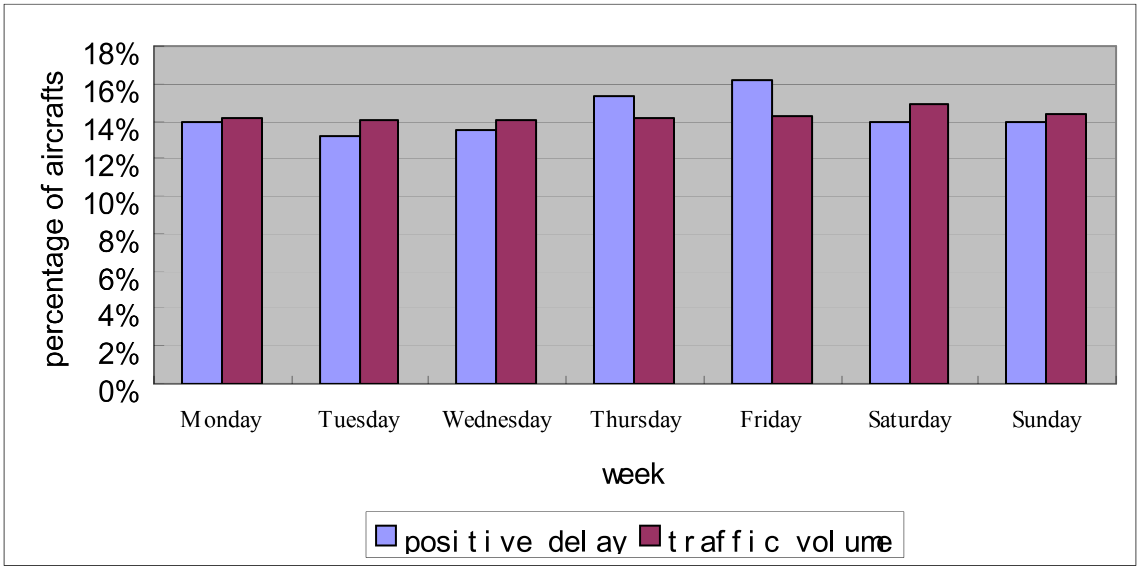

Figure 3.5 Delay distribution of arrivals as a function of day of week

Airports are more congested during certain times of day, and this may affect the delay distribution during a day. This distribution may also be dependent on the departure station or the arrival station. In Figure 3.6, the delay distribution during time of day are shown.

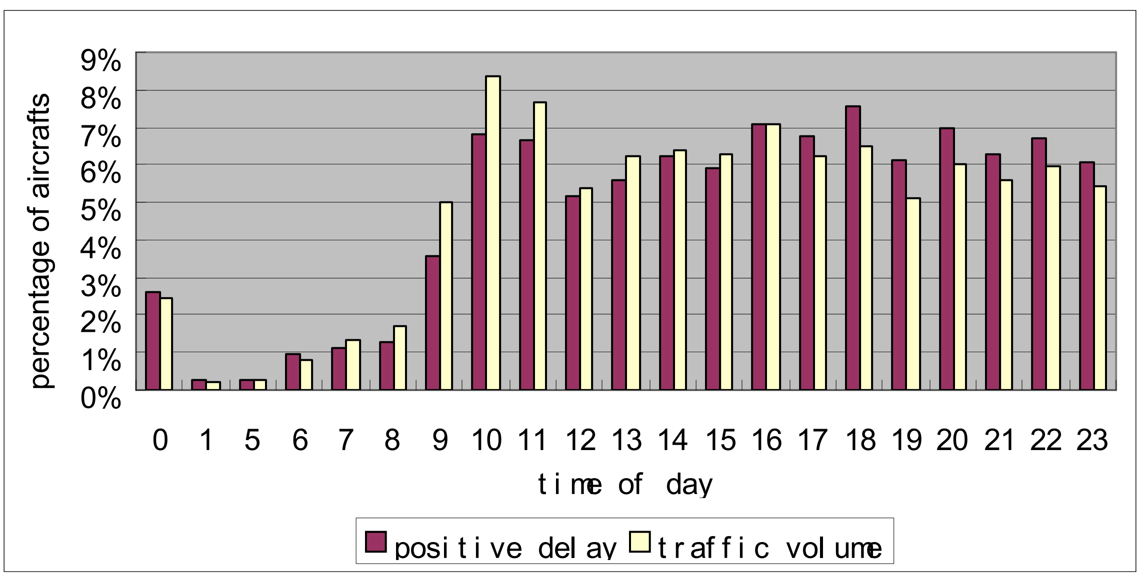

Figure 3.6 Delay distribution of arrivals as a function of time of day

Figure 3.6 shows the great variation in traffic volume in one day. From 1:00am to 5:00am, the volume is very low, while it peaks between 10:00am and 11:am. In the afternoon and evening, the volume still shows variations. Observe that although in the morning, when the volume is higher, the percentage of delayed aircrafts is lower than the percentage of traffic volume. This trend is reversed in the evening. The percentage of volume is lower than the percentage of delayed aircrafts. This can be due to many reasons. The traffic demand in the evening decreased, at the same time the capacity decreased. And it may be because of the origin airport. One possible explanation is the small variation in departures from day to day is not enough to reach a capacity threshold that will increase the number of delayed aircraft.

3.3.3 Arrival delay distribution according to flight characteristic

The delay distribution may depend on several characteristics of the flight. For example, a flight with a long scheduled airborne time may experience more variance in its actual airborne time than a flight with a short one.

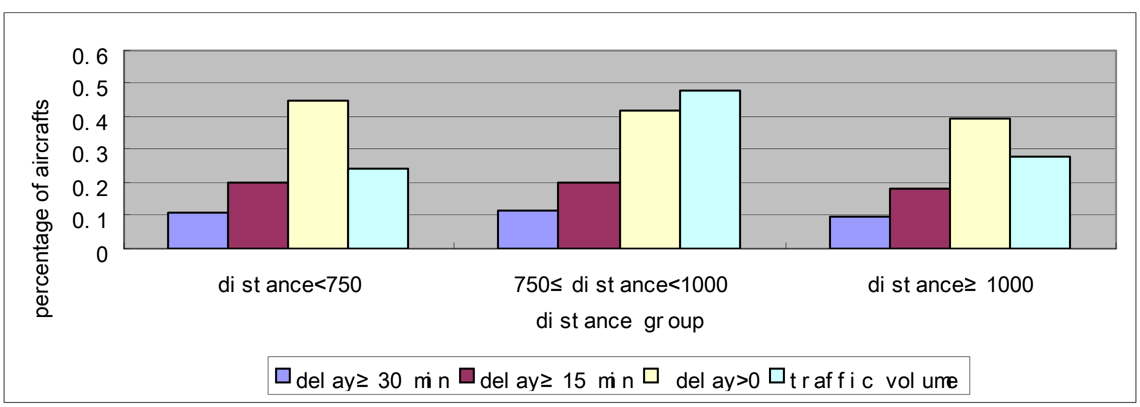

Figure 3.7 Delay distributions as a function of flight distance

Figure 3.7 shows the delay distribution as a function of flight distance. About 44.78% of the traffic volumes with flight distance <750 miles are positively delayed.

Among them, 19.73% are flights with delay ≥15, and 10.9% are flights with delay≥30. About 41.95% of the traffic volumes with flight distance from 750 to 1000 miles are with delay>0. Among them, 20.1% are flights with delay ≥15, and 11.5% are flights with delay≥30. About 39.4% of the traffic volumes with flight distance ≥ 1000 miles are with delay>0. Among them, 18% are flights with delay ≥15, and 9.9% are flights with delay≥30. The flights with flight distance <750 miles seem to be more probable to be positively delayed. The pattern of delays greater than 15 minutes is similar as the pattern of delays greater than 30 minutes.

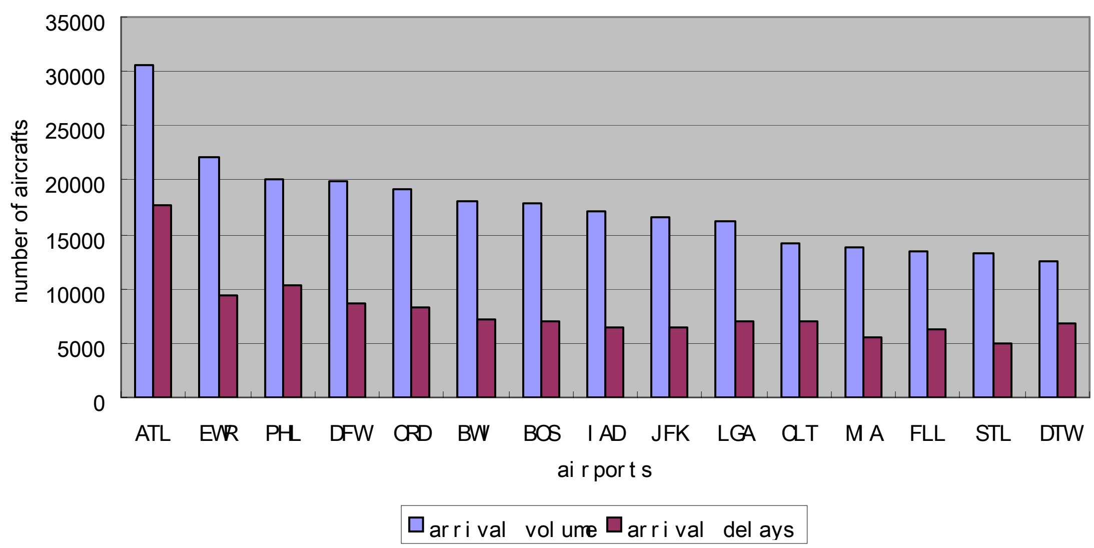

Figure 3.8 Arrival volume and arrival delays from top 15 airports

From Figure 3.8, the top 15 airports with the highest traffic volume to MCO (blue bar) and positively delayed flights (red bar) are shown. ATL shares the highest traffic volume in MCO, while the delay phenomenon is the most serious. 57.5% of arriving flights from ATL are delayed. Another airport needed to emphasize is DTW. Although the traffic volume in DTW ranks 15, 55.4% of the flights from DTW are with positive arrival delay. Also, at PHL and CLT, the positively delayed flights represent more than 50 percent of the whole volumes.

CHAPTER 4 AVERAGE DAILY DELAY MODEL

Based on the evidence on flight delays for the case of MCO described in Chapter 3, and on previous works in the literature analyzing this problem, in this chapter we develop statistical models to study airport congestion and delays.

This part was devoted to the analysis of departure and arrival delays of aircraft with the objective of detecting airport delay patterns and finding the contributing factors. To put the results in perspective, historical delay data for MCO from January 1 2002 to December 31 2003 were used. Causal factors for the delays related to aircraft, airline operations, change of procedures and traffic volume were identified.

4.1 Airport delay distribution and evaluation

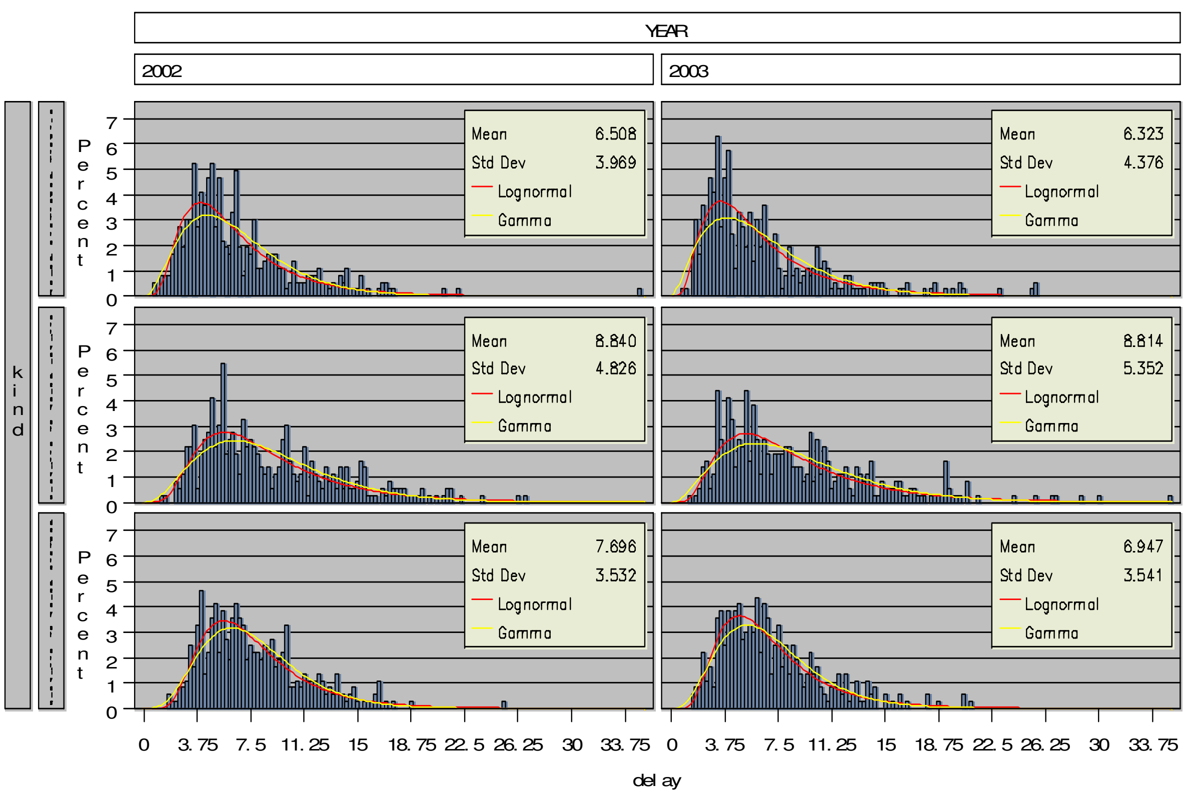

Figure 4.1 Delay distributions at MCO airport

Figure 4.1 shows trends in daily average arrival delay and departure delay at MCO, along with that of the departure delay of the arriving flights at MCO in 2002 and 2003. The vertical 3 figures are individually daily average departure delay of flights departing from MCO, daily average arrival delay of flights arriving at MCO and daily average departure delay of flights arriving at MCO. Lognormal distribution (in dark line) and Gamma (in light line) distribution are tried on these daily average delays of airports and Lognormal distribution is found to give a better fit. The calculations of these daily average delays are the same as the calculation of daily average arrival delay. Although daily average departure delay of arrivals decreases from 7.696 to 6.947 minutes, the daily average arrival delay and departure delay at MCO shows no change between 2002 and 2003.

Table 4.1 Correlation matrix of airport delay

	Arrival delay of arrivals at MCO	Departure delay of arrivals in other airports	Departure delay of departures at MCO
Arrival delay of arrivals at MCO	1.00	0.8726	0.8745
Departure delay of arrivals in other airports	0.8726	1.00	0.7691
Departure delay of departures at MCO	0.8745	0.7691	1.00

Given the correlation matrix among the airport delays, the arrival delay of arriving flight at MCO is highly related with its departure delay, which proves the analysis in section 3.3. And the arrival delay of arriving flights at MCO also shows high relation with the departure delay of departing flights at MCO. The arrivals and departures share the common facilities in the MCO airport, so the fluctuation in operational performance will definitely propagate each other. It is reasonable that the departure delay at other airports is highly related with the departure delay at MCO, since they are in the NAS and the delay phenomenon will spread out among them.

Table 4.2 2002-2003 MCO airport delay statistics

		2002	2003
Average daily departure delay	Mean(min)	6.508	6.323
	Std dev	3.969	4.376
Average daily arrival delay	Mean(min)	8.840	8.814
	Std dev	4.826	5.352
Annual departure delay for late flights (hr) (ADD)		7268.8	8369.2
Annual airborne time for departures at MCO (hr) (AATD)		160286.0	182276.8
Ratio of ADD/AATD		4.53%	4.59%
Annual arrival delay for late flights (hr) (AAD)		9929.8	11256.15
Annual airborne time for arrivals at MCO (hr) (AATA)		166197.7	189125.3
Ratio of AAD/AATA		5.97%	5.95%
Percentage of late departing flights		13.55%	13.23%
Percentage of late arriving flights		20.78%	19.85%

* Flights that leave/arrive at the gate more than fifteen minutes after the scheduled time shown in the carriers' CRS are considered late.

The items in Table 4.2 can be used as criteria to critique the airport operations based on performance. The percentage of late arriving flights at MCO is 20.78% and 19.85% in 2002 and 2003 respectively, which are relatively high values. It means that this airport may need additional capacity or operational improvements. The annual airborne time is calculated by the sum of flight time of departures or arrivals. We can see in Table 4.2 the annual airborne time of MCO is increasing quickly, and the annual arrival (or departure) delay is also increasing quickly. The ratio of this two can also be criteria to evaluate the airport delay. Average daily departure delay and arrival delay are another criteria, which can give more information about airport delay. Analysis on average daily arrival delay will be the focus in later sections.

4.2 Linear regression model of the average daily delay

We analyze the flight delay at the MCO airport in Orlando by estimating linear regression model of average daily delay that incorporates the effects of arrival demand, airport capacity, weather conditions in Orlando, seasonal effects. From the estimation results we are able to quantify some of the sources of delays from January 2002 to December 2003 and track changes in delays that are attributed to major causal factors.

4.2.1 Model description and variables

The statistical models are formulated based on data obtained from the Federal Aviation Administration (FAA) and the National Climatic Data Center (NCDC) as described in Chapter 3.

The multiple regression methodology was initially attempted. We assume that the errors are normally, identically, and independently distributed. Initial experimentation revealed that these assumptions do not hold. In particular we found that the errors are not normal. About the random error ε, the assumption that ε is normally distributed is the least restrictive when we apply regression analysis in practice.

After some experimentation, a model is developed for the average daily arrival delay at MCO in Table 4.1. The following describes general structure of a linear model including main effects and the interaction factors.

$d(t)= f(\theta,C(t), A(t), F(t), I(t), D(t), W(t), S(t), R(t), S(t)*W(t))+ v(t)$

where:

$d(t)$ is average arrival delay on day t;

$f(\cdot)$ is a model function;

$C(t)$ is maximum hourly capacity (including departure and arrival flight) in MCO;

$A(t)$ is a vector represents the arrival demand on day t in MCO;

$F(t)$ is the variables capturing different flight durations;

$I(t)$ and $D(T)$ are the space of inter-arrival time and its standard deviation;

$W(t)$ is a vector characterizing weather on day t;

$S(t)$ is a vector capturing seasonal influences;

$R(t)$ is a matrix capturing original airport regional variables;

$S(t)*W(t)$ is the interaction effect between seasonal effect and weather effect;

$v(t)$ is a stochastic error term.

In section 3.2, the arrival delay is described. For our delay metric, d(t), we used the average daily positive delay for all scheduled and completed flights from other airports to MCO airport in Orlando. It is the average of positive delay per flight per day. Flights that arrive early are assigned zero delay in the calculation.

The flight time variables show the influence of airborne time on airport delay. This effect is represented by four variables, which is respectively the percentages of arrivals with the flight time of 0- 1:59 hours, 2 hours to 2:59 hours, 3 hours – 3:59 hours, and more than 4 hours. They are created by the bins with approximately equal frequencies. So there are three factors in the model.

The space is calculated by the intervals between two consecutive arriving flights according to the scheduled time. When the space is smaller than 10 minutes (which is found to be a sensitive point to arrival delay in this dataset), it is assume 1, otherwise it is 0. The space variable we use in this model is calculated by the percentage of flights with the space smaller than 10 minutes.

Seasonal effects can be captured by 3 dummy variables as described before. The seasonal variables here are calculated by the percentage of flights that depart in each season.

The areas of the origin airports are divided into four parts that are southeast, southwest, northeast and northwest (appendix A). So this influence can be captured by 3 variables. And the percentage of daily flights from each site to MCO is calculated as a variable, which indicated the location effect on the flight.

4.2.2 Model Result

The model is introduced in Table 4.3 and 4.4.

Table 4.3 Model Fit Statistics for the linear model of average daily arrival delay

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	5220.68208	745.81173	39.07	<.0001
Error	722	13782.56230	19.08942		
Corrected Total	729	51.96			

Table 4.4 Model estimation for the linear model of average daily arrival delay

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.898758807	0.62127819	9.49	<.0001
Precipitation	0.040140903	0.00402798	9.97	<.0001
Wind	0.018784278	0.00646164	2.91	0.0038
Monday and Sunday	-1.604567713	0.42864250	-3.74	0.0002
Tuesday and Saturday	-2.683561807	0.42793584	-6.27	<.0001
Wednesday	2.012502691	0.52356437	-3.84	0.0001
Thursday and Friday	0.000000000			
Spring and Winter	2.657272124	0.39922837	6.66	<.0001
Summer	4.423730962	0.47084648	9.40	<.0001
Fall	0.000000000			

4.2.3 Model interpretation

The R square values for this linear model is only 0.2755, which is not satisfactory. And different transformations of dependent variable are tried including log, inverse, square, square root, exponential and standardized, but the R square did not improve. So some other statistics methods will be used to analyze the daily delay in the following section.

The model presented includes only 4 variables originally identified; the others were found to be statistically insignificant and eliminated.

In the model, delay is positively related to the precipitation, which means the higher rainfall will cause more flight delay. And the variable of wind contributes to the delay, which means the delay increases with higher wind speed.

As to the seasonal effect estimates, we find that delay increases during Summer relative to Fall. During spring and winter, the delay is higher than in fall. In summer, Orlando has much more heavy rainfall activity along with the hurricane that increases the delay. And one factor contributing to the pattern of seasonal effects is changes in upper air wind patterns throughout the year.

The results of the model shows significant weekly pattern. On Wednesday, Thursday and Friday, the average daily arrival delay shows higher value.

4.3 Analysis of Variance (ANOVA) on the average daily arrival delay

Analysis of variance (ANOVA) is used here to study the effects of one or more independent (predictor) variables on the dependent variable. Most commonly, ANOVA is used to test the equality of means by analyzing the total sum of squares (about the combined mean), which is partitioned into different components (due to model or due to random error).

Table 4.5 Tukey's Studentized Range (HSD) Test for daily arrival delay on week pattern

Week Comparison	Difference Between Means	95% Confidence Limits		
5 - 4	0.6788	-0.5863	1.9440	
5 - 7	2.0100	0.7449	3.2752	***
5 - 1	2.1348	0.8697	3.4000	***
5 - 3	2.3324	1.0702	3.5945	***
5 - 2	3.0037	1.7416	4.2659	***
5 - 6	3.0181	1.7529	4.2832	***
4 - 5	-0.6788	-1.9440	0.5863	
4 - 7	1.3312	0.0660	2.5963	***
4 - 1	1.4560	0.1908	2.7212	***
4 - 3	1.6535	0.3914	2.9157	***
4 - 2	2.3249	1.0627	3.5870	***
4 - 6	2.3392	1.0740	3.6044	***
7 - 5	-2.0100	-3.2752	-0.7449	***
7 - 4	-1.3312	-2.5963	-0.0660	***
7 - 1	0.1248	-1.1403	1.3900	
7 - 2	0.9937	-0.9072	2.8946	
7 - 3	0.3224	-0.9398	1.5845	
7 - 6	1.0080	-0.2571	2.2732	
1 - 5	-2.1348	-3.4000	-0.8697	***
1 - 4	-1.4560	-2.7212	-0.1908	***
1 - 7	-0.1248	-1.3900	1.1403	
1 - 3	0.1975	-1.0646	1.4597	
1 - 2	0.8689	-0.3933	2.1310	
1 - 6	0.8832	-0.3820	2.1484	
3 - 5	-2.3324	-3.5945	-1.0702	***

3 - 4	-1.6535	-2.9157	-0.3914	***
3 - 7	-0.3224	-1.5845	0.9398	
3 - 1	-0.1975	-1.4597	1.0646	
3 - 2	0.6713	-0.5878	1.9304	
3 - 6	0.6857	-0.5765	1.9478	
2 - 4	-2.3249	-3.5870	-1.0627	***
2 - 7	-0.9937	-2.2558	0.2685	
2 - 1	-0.8689	-2.1310	0.3933	
2 - 3	-0.6713	-1.9304	0.5878	
2 - 5	-3.0037	-4.9046	-1.1028	***
2 - 6	0.0143	-1.2478	1.2765	
6 - 5	-3.0181	-4.2832	-1.7529	***
6 - 4	-2.3392	-3.6044	-1.0740	***
6 - 7	-1.0080	-2.2732	0.2571	
6 - 1	-0.8832	-2.1484	0.3820	
6 - 3	-0.6857	-1.9478	0.5765	
6 - 2	-0.0143	-1.2765	1.2478	

** 1-7 represent Monday to Sunday individually；*** means showing significant difference with 95% confidence.

.

Table 4.6 Tukey's Studentized Range (HSD) Test for daily arrival delay on seasonal pattern

Season Comparison	Difference between means	95% Confidence Limits		
2 - 4	1.6876	0.4331	2.9421	***
2 - 1	2.9905	1.7429	4.2381	***
2 - 3	5.0435	3.7925	6.2946	***
4 - 2	-1.6876	-2.9421	-0.4331	***
4 - 1	1.3029	0.0484	2.5574	***
4 - 3	3.3559	2.0980	4.6138	***
1 - 2	-2.9905	-4.2381	-1.7429	***
1 - 4	-1.3029	-2.5574	-0.0484	***
1 - 3	2.0530	0.8020	3.3040	***
3 - 2	-5.0435	-6.2946	-3.7925	***
3 - 4	-3.3559	-4.6138	-2.0980	***
3 - 1	-2.0530	-3.3040	-0.8020	***

** 1-4 represent Spring to Winter individually.

The week differences are proved by F-test to be significant ($p<0.0001$). LSD test and TUKEY test both proved that the daily delay on Thursday and Friday are obviously higher than other weekdays. From Table 4.5, Tuesday and Saturday have the lowest daily delay in the week. This pattern should be related with the weekly schedule of the airport. At the same time, the season differences are proved by F-test to be significant ($p<0.0001$). From Table 4.6, LSD test and TUKEY test both proved that in summer the daily delay is obviously higher than other three seasons, and in fall the daily delay is obviously lower than other three seasons. In spring the daily delay is lower than in winter. This pattern should be related with several reasons. In Orlando summer is a rainy season. The thunder may cause the interrupt of the operations of the airport, which increase the delay of the flights that schedule arrival times are during or directly after the bad weather.

4.4 Using Proportional Odds Model to analysis the average daily arrival delay at MCO

4.4.1 Model description and variables

Logistic regression belongs to the group of regression methods for describing the relationship between explanatory variables and a discrete response variable. A logistic regression is proper to use when the dependent variable is categorized and can be applied to test the association between a dependent variable and the related potential factors, to rank the relative importance of independent variables, and to assess interaction effects.

To introduce the factors into logistic regression model and test their main effects on airport delay, the average daily arrival delay at MCO are identified, which are categorized into three groups. From the Table 4.7, the dependent variable (average daily arrival delay) can take on three values: $Y = 0$ for delay$<$5 min; $Y=1$ for delay $\geq$5 and $<$ 10 min; $Y=2$ for delay$\geq$10min. They are created by the bins with approximately equal frequencies.

Table 4.7 Quantiles of daily average arrival delay

Quantile	Estimate
100% Max	35.05098
99%	25.55390
95%	19.12648
90%	15.82271
75% Q3	11.67059
50% Median	7.52775
25% Q1	4.95496
10%	3.50041
5%	3.06098
1%	2.15493
0% Min	1.28016

Here we treat the arrival delay as a categorical outcome with three levels and keep the natural ordering presented in the data. There are usually three different ways of generalizing the logit model to handle ordered categories. We will use the Proportional Odds Model (Cumulative Logit Model). For this model, we have actually imposed the restriction that the regression parameters except the intercepts are the same for the two logit models. It implies that it doesn't make any difference how we categorize the dependent variable - the effects of the explanatory covariates are always the same. The results of modeling are showed below in Table 4.8, 4.9 and 4.10.

The independent variables are the same as the linear regression model in section 4.2. Considering the week pattern of the daily delay, the week is classified into five levels:

Monday and Sunday have no significant difference and are combined into one level. Tuesday and Saturday are combined into one level. Wednesday, Thursday and Friday are individually one level.

4.4.2 Model results

Table 4.8 Model estimation for logistic regression model of average daily arrival delay

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr>ChiSq
Intercept 3	1	-1.4073	0.1974	50.8256	<.0001
Intercept 2	1	0.8087	0.1927	17.6189	<.0001
precipitation	1	0.0264	0.00364	52.4524	<.0001
Monday and Sunday	1	0.4423	0.1935	5.2256	0.0223
Wednesday	1	0.4659	0.2355	3.9136	0.0479
Thursday	1	0.9280	0.2402	14.9235	0.0001
Friday	1	1.3301	0.2457	29.3148	<.0001
spring	1	-0.2356	0.2026	1.3520	0.2449
summer	1	0.3872	0.2161	3.2104	0.0732
fall	1	-1.4521	0.2113	47.2132	<.0001

Table 4.9 Odds Ratio Estimates for logistic regression model of average daily arrival delay

Effect	Point estimate	95% Wald Confidence Limits	
Precipitation	1.027	1.019	1.034
Monday and Sunday vs Tuesday and Saturday	1.556	1.065	2.274
Wednesday vs Tuesday and Saturday	1.593	1.004	2.528
Thursday vs Tuesday and Saturday	2.530	1.580	4.051
Friday vs Tuesday and Saturday	3.781	2.336	6.120
spring vs winter	0.790	0.531	1.175
Summer vs winter	1.473	0.964	2.250
Fall vs winter	0.234	0.155	0.354

Table 4.10 Model Fit Statistics for logistic regression model of average daily arrival delay

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.0	Somers' D	0.504
Percent Discordant	23.6	Gamma	0.516
Percent Tied	2.4	Tau-a	0.330
Pairs	174384	c	0.752

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
5.5864	7	0.5888

4.4.3 Model interpretation

From table 4.10, test statistic for the Proportional Odds Assumption is 5.5864with the DF of 7, so the p value is 0.5888. The high p-value is desirable. For this problem, we find no reason to reject the proportional odds model.

Adjusting for other variables, the odds of having a higher arrival delay in spring will be 0.79 times of the odds in winter, the odds of having a higher arrival delay in summer will be 1.473 times of the odds in winter, and the odds of having a higher arrival delay in fall will be 0.234 times of the odds in winter.

On Monday and Sunday the odds of having a higher arrival delay will be 1.556 times of the odds on Tuesday and Saturday, on Wednesday the odds of having a higher arrival delay will be 1.593 times of the odds on Tuesday and Saturday, on Thursday the odds of having a higher arrival delay will be 2.530 times of the odds in Tuesday and Saturday, and on Friday the odds of having a higher arrival delay will be 1.556 times of the odds in Tuesday and Saturday. From the odds ratio, on Thursday and Friday the airport is showed to have the higher probability to have delay more than 10 minutes.

For each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0264*10)-1= 30.2%. There are no significant interactions.

Compared with linear regression model, the multiple logistic regression model shows the same seasonal pattern and weekly pattern. But the logistic regression model shows a better fit of the dataset. The daily delay on Thursday and Friday are obviously higher than other weekdays. Tuesday and Saturday have the lowest daily delay in the week. In summer the daily delay is obviously higher than other three seasons, and in fall the daily delay is obviously lower than other three seasons. In spring the daily delay is lower than in winter. At the same time the variable of precipitation contributes to the delay.

From the results we find that all significant factors are uncontrollable, and the average daily delay is related to the weather conditions.

4.5 Using neural network to analyze the average daily arrival delay at MCO

4.5.1 A brief review of methodology

Artificial neural networks are alternative computation techniques that can be applied to solve categorical analysis problems. In this section we describe multi-layer perceptron (MLP) and Radial basis function (RBF) neural network that are most commonly used neural network architectures.

The MLP network is one of the most popular neural network architectures that fit a wide range of applications such as forecasting, process modeling, and pattern discrimination and classification. MLPs are feed-forward neural networks trained with the standard back-propagation algorithm. They are supervised networks so they require a desired response to be trained.

A radial basis function (RBF) network is a feed forward network with a single hidden layer for which the 'combination function' is more complex and is based on a distance function (referred to as width) between the input and the weight vector. Ordinary RBF (ORBF) networks using radial combination function and exponential activation function are universal approximators in theory (Powell, 1987), but in practice they are often ineffective estimators of the multivariate function. Due to the localized effect the ORBF neural networks often require an enormous number of hidden units to avoid an unnecessarily bumpy fit. To avoid the pitfalls of ORBF networks, softmax activation function may be used. It essentially normalizes the exponential activations of all hidden units to sum to one. This type of network is called a "normalized RBF" or NRBF network. NRBF is used in this section.

4.5.2 Model results and conclusions

The models in this section are formulated based on the same data as used in the linear and logistic regression models. The sample size is 730. The difference is the data is partitioned into two parts, 60% for training model and 40% for validation.

8 MLP and NRBF neural network models having a range (2 to 5) of hidden nodes are compared. The result from these neural networks shows that the MLP network with 4 hidden nodes performs best among the models in their respective architectures. At the same time, one logistic regression model is employed to compare with the best MLP network.

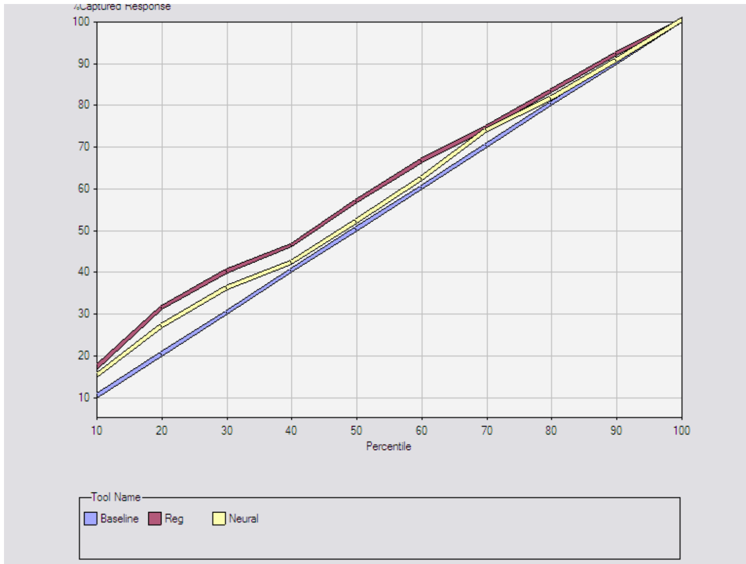

Figure 4.2 The captured response lift plots for models of the daily delay

The captured response lift plots for models of the daily delay are showed in Figure 4.2. The best models were identified through the lift plot having cumulative percentage of captured response in the validation dataset on vertical axis. The higher a curve from the baseline curve the better is the performance of the corresponding model. It may be noted that the logistic regression model has its captured response percentage higher than the MLP model within first seven deciles (deciles = 10 percentiles). Since the 3-level target is classified by about same frequency, the three levels (lower delay(<5minute), medium delay (5-10 minute) and higher delay(≥10 minute)) will be individually about 30-40% of the whole data and therefore it is decided to evaluate the model performances within first four deciles (deciles = 10 percentiles). So we can say that the logistic regression model performs better that neural network models for this daily delay data.

4.6 Conclusions

Not all arrivals can occur when they are scheduled to, then airport congestion happens. The delay distribution of the airport can make it easier to understand the airport delay. The assessment of an airport's schedule performance is also discussed. Finally, Multivariate regression, ANOVA, Neural networks and Logistic regression are used to detect the pattern of airport arrival delay.

The results of our research show that the arrival delay is highly related to the delay at the origin. The airport arrival delay can also be used to evaluate the airport delay. The airport arrival delay is found to show seasonal and weekly patterns, which are related to the schedule performance. The precipitation and wind speed are also found contributors of airport arrival delay. The capacity of the airport is not found to be significant. This may indicate that the capacity constraint is not a determinant variable in the delay problem at MCO.

CHAPTER 5 SINGLE FLIGHT ARRIVAL DELAY MODELS

From the Airline On-Time Performance Data, the airline's on-time performance in the Orlando International Airport (MCO) —the proportion of flights arriving within 15 minutes after scheduled time, for 2004 was 68.20 percent, decreased from 69.09 percent in 2003, while the traffic volume increase from 92357 to 107871 per year. These delays are frustrating to air travelers and costly to airlines. They are also the concern in this research. What is the pattern of the delay? What is the contribution of various causal factors? Can we predict the delay with the flight schedule?

Historical data exists to describe nearly all of the random events and variables involved. However, no simple formula exists that allows the scheduler to measure their complex interaction. The key to solving the airline scheduling problem is to recognize the random processes involved and make scheduling and policy decisions that minimize the risk of delays. To allow the scheduler to test a variety of scheduling strategies and operations policies that might impact schedule performance, we focus on all the delayed flights under schedule conditions to minimize their interaction.

This research analyzes different factors that affect flight delays and flights with high delay. Logistic regression is used to analyze how airport factors, airline factors and weather conditions influence delay at MCO. Although sophisticated simulation models can be used to predict delay, they are not well suited to our goal of assessing the sensitivity of delay to individual flights in the schedule, since every question requires a new simulation run. Instead, we employ a set of logistic regression models that both predicts delay and clearly reveals the sensitivity of factors to individual flight delay. The multiple logistic regression is tried, with the delay divided into 3 levels: 0 minutes≤ delay<15 minutes, 15 minutes≤delay<30 minutes, and delay≥30 minutes. But the results show not satisfactory. So two binary logistic regression models are used in this chapter to illustrate the pattern of delay.

5.1 Delay model on the flights with delay ≥0

5.1.1 General

The flights that leave the gate more than fifteen minutes after the scheduled time shown in the carriers' computerized reservations systems (CRS) are considered "late" while all other flights are recorded as "on-time". Similar to origin airports, flights at destination airports are defined as "on-time" if they arrived at the gate within fifteen minutes of the scheduled time shown in the carriers' CRS, while all remaining flights are defined as "late". The delayed flights considered here are the flights with delay equal to or more than 15 minute.

To introduce the factors into a statistical model and test their main effects on delay of late flights, only the on-time flights and delayed flights from 2002 to2003 (shown in Table 3.1) are identified, which are categorized into two groups: no delay (0≤arrival delay<15), delayed flight (15≤arrival delay).

Those factors include information of individual flights, as well as the corresponding airport conditions and weather conditions. So that the dependent variable Y (individual flights) here takes on two values: Y = 0 for no-delayed flights (0≤arrival delay<15), Y = 1 for late flights (15≤arrival delay). From Table 5.1, the no-delayed flights and late flights respectively represent 61.54% and 38.46% of the whole data set.

Table 5.1 Sample size Sample size for flight with delay ≥0

Period	Number of flights		
	no-delayed flights	late flights	Total
2002	21701	13874	35575
2003	24701	15123	39824
Total	46402	28997	75399

Binary logistic regression is proper to use here when the dependent is a dichotomy (an event happened or not) and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects.

The model was estimated on a data set consisting of all the individual domestic arriving flights with delay ≥0 at MCO. The data for the non-stop flights on scheduled service by certificated carriers to MCO were obtained from the Airline On-Time Performance Data on the Bureau of Transportation Systems website from 01/01/2002 through 12/31/2003 excluding the cancelled and diverted flights.

The factors are introduced as the same as delay model in section 2.1. The significant independent variables in the model are introduced in Table 5.2.

Table 5.2 Definition of independent variables for delay model with delay≥0

Parameter	Definition
Thursday and Friday	The flight takes place on Thursday and Friday
Summer	The flight takes place in summer
Winter	The flight takes place in winter
Fall	The flight takes place in fall
Evening	The flight takes place between 5pm to 11:59pm
Afternoon	The flight takes place between 12pm to 4:59pm
Distance between 750-1000	The flight distance is in between 750 and 1000 miles
Distance > 1000	The flight distance is larger than 1000 miles
Precipitation	Hundredth of inches of the precipitation per day
log space	The log transform of the space

5.1.2. Model results

The final logit model is presented in Tables 5.3 and 5.4 and 5.5.

Table 5.3 Model estimation for delay model on the flights with delay≥0

Parameter		Estimate	StandardError	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0827	0.0287	1423.1896	<.0001
Winter	1	0.1432	0.0215	44.5419	<.0001
Fall	1	-0.2595	0.0233	123.7075	<.0001
Summer	1	0.2191	0.0211	107.7185	<.0001
Evening	1	0.7179	0.0210	1166.2651	<.0001
Afternoon	1	0.2889	0.0226	163.9135	<.0001
Thursday and Friday		0.1172	0.0164	50.8101	<.0001
Distance > 1000		-0.0338	0.0214	2.4898	0.1146
Distance between 750 and 1000		0.1921	0.0182	111.4865	<.0001
Precipitation		0.0226	0.00111	414.5916	<.0001

log space	-0.0198	0.00905	4.7965	0.0285

Table 5.4 Odds Ratio Estimates for delay model on the flights with delay≥0

Effect	Point Estimate	95%Wald Confidence Limits	
Winter vs spring	1.154	1.106	1.203
Fall vs spring	0.771	0.737	0.808
Summer vs spring	1.245	1.194	1.297
Evening vs moring	2.050	1.967	2.136
Afternoon vs moring	1.335	1.277	1.395
Thursday and Friday vs other weekdays	1.124	1.089	1.161
Distance>1000vs distance<750	0.967	0.927	1.008
Distance in[750,1000]vs distance<750	1.212	1.169	1.256
Precipitation	1.023	1.021	1.025
log space	0.980	0.963	0.998

Table 5.5 Model Fit Statistics for delay model on the flights with delay≥0

Criterion	Intercept Only	intercept & Covariates
AIC	99612.098	96759.608
SC	99621.319	96861.041
-2 Log L	99610.098	96737.608

Testing Global Null Hypothesis: BETA=0			
Likelihood Ratio	2872.4895	10	<.0001
Score	2785.6663	10	<.0001
ald	2553.7276	10	<.0001
Likelihood Ratio	2872.4895	10	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.9	Somers' D	0.224
Percent Discordant	38.4	Gamma	0.226
Percent Tied	0.7	Tau-a	0.106

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
17.4662	8	0.1226

5.1.3 Model interpretation

From Table 5.5, Hosmer and Lemeshow Goodness-of-Fit Test statistic is 17.4662 with the DF of 8. The resulting p value of 0.1226, shown in Table 5.5, suggests that the model fits well. We find no reason to reject the odds model at 5% confidence level.

Four factors including season influences, time influences, distance influences, precipitation and space show significant association with the likelihood of flights of delay.

The ‘odds ratio’ column in Table 5.4 are obtained from the parameter estimates. For the dummy variable ‘summer’, which indicates the flight takes place in summer, the odds ratio is 1.245. The odds ratio of 1.245 tells us that the predicted odds of ‘delay flights’ for summer are 1.245 times the odds for other seasons. In other words, the predicted odds of delay are about 24.5% higher when the season is summer. The dummy variable ‘winter’ indicates the flight takes place in winter, the odds ratio is 1.154. This implies that the predicted odds of delay are about 15.4% higher when the season is winter. The dummy variable ‘fall’ indicates the flight takes place in fall, the odds ratio is 0.771. This implies that the predicted odds of delay are about 23% lower in winter than other seasons.

The odds ratios for the arrival time show the relative ratios of high delay between different times (morning, afternoon, an evening) for each flight. Compared to morning, the odds of delay in the afternoon could be 1.335 times higher and the odds in the evening could be 2.05 times higher. At evening the flights are much more likely to be delayed than in morning. The results testified that the time would definitely contribute to delay.

The odds ratio of the distance variables is interesting. For each flight with the flight distance between 750 to 1000 miles, the odds of having arrival delay will increase by 1.212-1=21.2%. While for the flight with the flight distance larger than 1000 miles, the odds of having arrival delay will decrease 1-0.967=3.3% than the flight with the flight distance less than 750 miles.

The odds ratio for the precipitation shows that for each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0226*10)-1= 25.3%.

The odds ratio for the log space shows that as the space increases, the probability of the flights being delayed will decrease. Since the space is the inter-arrival time for two successive flights, when the space increases, the service time for each flight will increase, so that the efficiency of the operation will decline.

5.2 Delay model of the flights with the delay≥15 minutes

In this section a data mining approach is presented to detect the pattern of the delayed (late) flights (with the delay≥15 minutes), which separate low-delay flights from high-delay flights. The formation and structure of the dataset used for the analysis are discussed in detail as follow.

5.2.1 Methodology

To introduce the factors into statistical model and test their main effects on the extent of flights delay, the late arriving flights at MCO (with arrival delay≥15 minute) are identified, which are categorized into two groups: low delay and high delay.

Those factors include information of late flights, as well as the corresponding airport conditions. So that the dependent variable Y (late flights) here takes on two values: Y = 0 for low-delayed flights (with 15-30 minutes delayed), and Y = 1 for high- delayed flights (with more than 30 minutes delayed). When the flight is delayed no more than 30 minute, it is considered low-delayed, otherwise high-delayed. From Table 5.6, the low-delayed flights and high-delayed flights respectively represent 46.49% and 53.51% of the flights with delay more than or equal to 15 minutes.

Table 5.6 Sample size for flight with delay ≥15 minutes

Period	Number of flights		
	Low delay	High delay	Total
2002	6663	7211	13874
2003	6819	8304	15123
Total	13482	15515	28997

Linear regression model is tried not to be satisfactory here, because the R square is very small and the error is not normal. Binary logistic regression is proper to use here when the dependent is a dichotomy (an event happened or not) and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects.

5.2.2 Model results

Table 5.7 shows the significant independent variables in the model and the definitions. The final logit model is presented in Tables 5.8 and 5.9 and 5.10.

Table 5.7 Definition of independent variables for delay model with delay≥15

Parameter	Definition
West	The origin airport is in west area
Summer	The flight takes place in summer
Winter	The flight takes place in winter
Evening	Schedule arrival time later than 4:59pm
Afternoon	Schedule arrival time from 12pm to 4:59pm
Distance between 750 and 1000	The flight distance is in [750, 1000]
Distance greater than 1000 mile	The flight distance is greater than 1000 miles
Distance greater than 750 mile	The flight distance is greater than 7500 miles
Distance >750mile *evening	The product of evening and fall
Distance >750mile *afternoon	The product of evening and weekend
Precipitation	Hundredth of inches of the precipitation per day

Table 5.8 Model estimation for delay model with delay≥15

Parameter	Estimate	StandardError	Wald Chi-Square	Pr > ChiSq
Intercept	-0.2409	0.0533	20.4052	<.0001
Winter	0.1032	0.0335	9.4638	0.0021
Fall	-0.0670	0.0378	3.1416	0.0763
Summer	0.1987	0.0327	36.9306	<.0001
West	-0.3630	0.0631	33.0936	<.0001
Evening	0.3440	0.1257	7.4913	0.0062
Afternoon	0.3814	0.1288	8.7765	0.0031
Wednesday	-0.1068	0.0349	9.3737	0.0022
Tuesday	-0.0746	0.0370	4.0543	0.0441
Distance greater than 1000 mile	0.1475	0.0593	6.1855	0.0129
Distance between 750 and 1000	0.0721	0.0559	1.6610	0.1975
Precipitation	0.00775	0.00101	58.6159	<.0001
Distance >750mile *evening	0.1138	0.0698	2.6574	0.1031
Distance >750mile *afternoon	-0.1053	0.0707	2.2147	0.1367

Table 5.9 Odds Ratio Estimates for delay model with delay≥15

Effect	Point Estimate	95%Wald Confidence Limits	
Winter vs spring	1.109	1.038	1.184
Fall vs spring	0.935	0.868	1.007
Summer vs spring	1.220	1.144	1.301
West vs others	0.696	0.615	0.787
Wednesday vs others	0.899	0.839	0.962
Tuesday vs others	0.928	0.863	0.998
Distance ≥1000 vs distance <750	1.159	1.032	1.302
Distance in[750,1000]vs distance <750	1.075	0.963	1.199
Precipitation	1.008	1.006	1.010

Table 5.10 Model Fit Statistics for delay model with delay≥15

Criterion	Intercept Only	intercept & Covariates
AIC	39796.695	39262.704
SC	39804.964	39378.464
-2 Log L	39794.695	39234.704

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	559.9913	13	<.0001
Score	551.6177	13	<.0001
Wald	539.5727	13	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	56.9	Somers' D	0.159
Percent Discordant	41.0	Gamma	0.163
Percent Tied	2.2	Tau-a	0.079
Pairs	206482128	c	0.580

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.3118	8	0.6123

5.2.3 Model interpretation

Output of the binary logistic regression includes model estimation and odds ratio estimate for significant independent variables and the model fit statistics.

There are some summary statistics to measure the goodness of fit for the regression model. That is a single number that represents a model fit. From Table 5.10, Hosmer and Lemeshow Goodness-of-Fit Test statistic is 6.3118 with the DF of 8. Hosmer and Lemeshow statistic is calculated in the following way. Based on the estimated model, predicted probabilities are generated for all observations. These are sorted by size, and then grouped into approximately 10 intervals. Within each interval the expected frequency is obtained by adding up the predicted probabilities. Expected frequencies are compared with observed frequencies by the conventional Pearson chi-square statistic. We find no reason to reject the odds model at 5% confidence level. The resulting p value of 0.6123, shown in Table 5.8, suggests that the model fits well.

Four factors including season influences, time influences, week influences, distance influences, regional influences and precipitation show significant association with the likelihood of flights of high delay.

Let's look at the numbers in the 'odds ratio' column in table 5.9, which are obtained from the parameter estimates. For the dummy variable 'summer', which indicates the flight takes place in summer, the odds ratio is 1.220. The odds ratio of 1.220 tells us that the predicted odds of 'high delay' for summer are 1.22 times the odds for other seasons. In other words, the predicted odds of high delay are about 22% higher when the season is summer. The dummy variable 'winter' indicates the flight takes place in winter, the odds ratio is 1.109. This implies that the predicted odds of high delay are about 10.9% higher when the season is winter. The dummy variable 'fall' indicates the flight takes place in fall, the odds ratio is 0.935. This implies that the predicted odds of high delay are about

6.5% lower when the season is fall.

The odds ratios for the arrival time show the relative ratios of high delay between different times (morning, afternoon, an evening) for each flight. It shows that there is a clear association between the ratios of high delay and the arrival time. Compared to morning, the odds of high delay in the evening could be $e^{0.4578}$=1.58 times higher and the odds in the afternoon could be $e^{0.2762}$=1.318 times higher. This means as that latter time periods in the day experience higher delay ratios. The results testified that the time would definitely contribute to high delay.

Adjusting for other variables, the odds of having a higher arrival delay when a flight departs from west area is $e^{-0.3630}$=0.695 times of the odds in other areas. That means the odds of having more arrival delay decreases by 1-0.695= 30.5% when the flight takes off from the west area.

For each flight with the flight distance between 750 to 1000 miles, the odds of having a higher arrival delay will increase by 1.075-1=7.5%. While for the flights with the flight distance more than 1000 miles, the odds of having a higher arrival delay will increase by 1.159-1=15.9%. The odds ratios show that the flights with long flight distance will have higher odds of high delay.

After confirming the main effect model, the next regression analysis is to explore the possible significant interactions between these factors. It is found that there is one interaction factors associated with high delay including: evening and distance >750mile (P-value = 0.1031), afternoon and distance >750mile (P-value = 0.1367).

The results confirm that the effects of distance are different between daytime. For the flight distance more than 750 miles, in evening high delay is more likely to occur than morning; while in afternoon high delay is less likely to occur than morning; for other distance groups, the difference between weekdays is not apparent.

5.3 Conclusions

Flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. And there exist some pattern of flight delay due to the schedule performance and airline itself. The results extracted from the case study on Orlando International Airport (MCO) can help to understand better the phenomenon.

From the delay models on the flights with delay≥0 and delay≥15min, the individual flight arrival delay is found to show seasonal pattern, weekly pattern, which corresponds to the pattern of the airport arrival delay. And the precipitation is also found contributors of airport arrival delay. The wind speed at MCO is found to influence the airport arrival delay, but is not so significant to the individual flight arrival delay. The capacity of airport is also found to be not significant to single flight delay.

From the delay models of the flights with delay≥0, the odds ratio of the distance variables is interesting. For each flight with the flight distance between 750 to 1000 miles, the odds of having arrival delay will increase by 21.2%. While for the flight with the flight distance larger than 1000 miles, the odds of having arrival delay will decrease 3.3% than the flight with the flight distance less than 750 miles. Another variable needed to mentioned is the space of successive flights. The odds ratio for the space shows that as the space increases, the probability of the flights being delayed will decrease. Since the space is the inter-arrival time for two successive flights, when the space increases, the service time for each flight will increase, so that the efficiency of the operation will decline.

From the delay models of the flights with delay≥15, the flights with long flight distance will have higher odds of high delay. The results testified that the time would definitely contribute to high delay, that latter time periods in the day experience higher delay ratios. Adjusting for other variables, the odds of having more arrival delay decreases by 30.5% when the flight takes off from the west area. The results also confirm that the effects of distance are different between daytime. For the flight distance more than 750 miles, in the evening high delay is more likely to occur than the morning; while in the afternoon high delay is less likely to occur than the morning; for other distance groups, the difference between weekdays is not apparent.

Capacity increase is not necessarily a solution to airport congestion, in a context of rapidly growing demand for air services. Another characteristic of air congestion that we illustrate empirically is that a flight delay is not necessarily more costly during a peak period, but depends on the impacts generated on subsequent flights. In response to single flight delay predictions and reason for these delays that are generated by the model, which can give indications for the appropriate recovery actions to recover/avoid these delays.

CHAPTER 6 ANALYSIS ON THE DELAY DUE TO MCO

Analysis of the airline's performance data shows that some delays can be attributed to operating procedures. An example is originating delay. That is, the first flight segment of the day for some lines of flying typically depart late. Originating delay concerns the airlines because it can impact the entire line of flying. The results of our research show that the arrival delay is highly related to the origin delay.

While it was possible to calculate the immediate impact of originating delays, it is not possible to calculate their impact on the cumulative delay. If a late originating aircraft has no slack in its down line schedule, it will continue to be late. If that aircraft enters a connecting bank, it can pass its lateness on to other aircraft. So here we purify only the arrival delay at MCO, excluding the flights with originating delay >0. The model will make it possible to see the pattern of the aircraft arrival delay. Of course, the result can range form insignificant to significant depending on the status of other control factors.

6.1 A brief review of methodology

The analytic goal is to predict the flight delay from the schedule information. Below is the description of the predictors. In this project, logistic regression model, tree model, and neural network will be used to fit the single flight delay and then compare them on the basis of an independent test sample. The delayed flights considered here are the arriving flights with the delay equal to or more than 1 minute. We only consider the delayed flights that depart before schedule time or on time, so that we can make sure these arrival-delayed flights are not departure-delayed.

To introduce the factors into statistical model and test their main effects on delay of flights, the delayed flights from other airports to MCO are identified, which are categorized into two groups: low delay and high delay.

The flights that arrive the gate more than fifteen minutes after the scheduled time are considered “high delay” while other flights with delayed time less than 15 minutes are recorded as “low delay”.

Those factors include information of delayed flights, as well as the corresponding airport conditions. So that the dependent variable Y (delayed flights) here takes on two values: Y = 0 for low-delayed flights (with 1-15 minutes delayed), and Y = 1 for high-delayed flights (with more than 15 minutes delayed). When the flight is delayed no more than 15 minute, it is considered low-delayed, otherwise high-delayed. From Table 6.1, the low-delayed flights and high-delayed flights respectively represent 79.54% and 20.46% of the whole data set.

Table 6.1 Sample size used for flight delay analyses

Period	Number of flights		
	Low delay	High delay	Total
2002	8070	2145	10215
2003	11139	2796	13935
Total	19209	4941	24150

6.2 Modeling results and analysis

6.2.1 Classification of delay based on logistic regression

Table 6.2, 6.3, 6.4 list the Maximum Likelihood Estimates and odds ratios properly adjusting other factors for significant independent variables, where the highest levels of independent variables are considered as the default levels. Table 6.5 lists the model fit statistics. The following sections document the interpretation of the regression results.

Table 6.2 Significant variables for Logistic regression model

Effect	DF	wald chi-square	P value	definition
Crs_time	3	67.14	<0.0001	Schedule arrival time of each flight
distance		76.44	<0.0001	Flight distance of each flight
Season		51.96	<0.0001	The season when the flights depart
sitecode		65.12	<0.0001	The area of the original airport
preciption		265.14	<0.0001	The number of rainfall per day
wind		9.39	0.0022	The wind speed per day

Table 6.3 Analysis of Maximum Likelihood Estimates for Logistic regression model

parameter	DF	Estimate	Stan error	wald chi-square	P value	definition
Intercept	1	-1.9133	0.0754	644.5	<0.0001	
Crs_time1	1	-0.4369	0.0744	34.51	<0.0001	Schedule arrival time of 7am to 8:59am
Crs_time2	1	0.0616	0.0397	2.40	0.1211	Schedule arrival time of 9am to 2:59 pm
Crs_time3	1	0.3062	0.0399	58.86	<0.0001	Schedule arrival time of 3pm to 8:59pm
Distance1	1	-0.3287	0.0404	66.20	<0.0001	Flight distance of 0 to 750 miles
Distance2	1	0.2349	0.0332	50.08	<0.0001	Flight distance of 750 to 1000 miles
Season1	1	-0.0437	0.0385	1.28	0.2570	Spring
Season2	1	0.2578	0.0381	45.72	<0.0001	Summer
Season3	1	-0.1734	0.0395	19.32	<0.0001	Fall
Sitecode1	1	-0.0651	0.0610	1.14	0.2861	South area
Sitecode2	1	0.2852	0.0399	51.10	<0.0001	East area
Sitecode3	1	-0.0899	0.0471	3.64	0.0565	Central area
preciption	1	0.0392	0.00241	265.14	<0.0001	The hundred inches of rainfall per day
wind	1	0.00259	0.000845	9.39	0.0022	The wind speed per day

Table 6.4 Odds Ratio Estimates for Logistic regression model

variables		Point estimate
Crs_time1	7am to 8:59am vs later than 9pm	0.603
Crs_time2	9am to 2:59pm vs later than 9pm	0.992
Crs_time3	3pm to 8:59pm vs later than 9pm	1.267
Distance1	Less than 750 miles vs longer than 1000 miles	0.655
Distance2	750 miles to 1000 miles vs longer than 1000 miles	1.151
Season1	Spring vs winter	0.997
Season2	Summer vs winter	1.348
Season3	Fall vs winter	0.876
Sitecode1	South vs west	1.067
Sitecode2	East vs west	1.515
Sitecode3	Central vs west	1.041
preciption		1.040
wind		1.003

Table 6.5 Hosmer and Lemeshow Goodness-of-Fit Test for Logistic regression model

Chi-Square	DF	Pr > ChiSq
9.6248	8	0.2924

Output of the binary logistic regression includes model estimation and odds ratio estimate for significant independent variables and the model fit statistics. There are some summary statistics to measure the goodness of fit for the regression model. That is a single number that represents a model fit. From Table 6.5, Hosmer and Lemeshow Goodness-of-Fit Test statistic is 9.6248 with the DF of 8. Hosmer and Lemeshow statistic is calculated in the following way. Based on the estimated model, predicted probabilities are generated for all observations. These are sorted by size, and then grouped into approximately 10 intervals. Within each interval the expected frequency is obtained by adding up the predicted probabilities. Expected frequencies are compared with observed frequencies by the conventional Pearson chi-square statistic. We find no reason to reject the odds model at 5% confidence level. The resulting p value of 0.2924, shown in Table 6.5, suggests that the model fits well.

Four factors including season influences, time influences, week influences, distance influences, regional influences and precipitation show significant association with the likelihood of flights of high delay.

Let's look at the numbers in the 'odds ratio' column in Table 6.4, which are obtained from the parameter estimates. For the dummy variable 'summer', which indicates the flight takes place in summer, the odds ratio is 1.348. The odds ratio of 1.348 tells us that the predicted odds of 'high delay' for summer are 1.348 times the odds for the winter season. In other words, the predicted odds of high delay are about 34.8% higher when the season is summer. The dummy variable 'spring' indicates the flight takes place in spring, with the odds ratio is 0.997. This implies that the predicted odds of high delay in spring are about the same as the season of winter. The dummy variable 'fall' indicates the flight takes place in fall, with the odds ratio is 0.876. This implies that in fall the predicted odds of high delay are about 12.4% lower than when the season is fall.

The odds ratios for the arrival time show the relative ratios of high delay between different times (7am to 8:59am, 9am to 2:59pm, 3pm to 8:59pm, 9pm to 11pm) for each flight. It shows that there is a clear association between the ratios of high delay and the arrival time. Compared to the time of 9pm to 11pm, the odds of high delay in the 7am to 9am could be 0.603 times, the odds in the 9am to 3pm could be 0.992 times, and the odds in the 3pm to 9pm could be 1.267 times. This means that in 7am to 9am there is less odds of high delay than in 9pm to 11pm, in 9am to 3pm the odds of high delay has no big difference from in 9pm to 11pm, while in 3pm to 9pm the odds of high delay will be 26.7% higher than in 9pm to 11pm.

Adjusting for other variables, the odds of having a higher arrival delay when a flight departs from south area is 1.067 times of the odds in west areas. The odds of having a higher arrival delay when a flight departs from east area is 1.515 times of the odds in west areas. The odds of having a higher arrival delay when a flight departs from central area are 1.041 times of the odds in west areas. That means when the flight takes off from the east area, the odds of having higher arrival delay will be about 45-50% higher than other areas.

For each flight with the flight distance between 750 to 1000 miles, the odds of having a higher arrival delay will increase by 1.151-1=15.1%. While for the flights with the flight distance less than 750 miles, the odds of having a higher arrival delay will decrease by 1-0.655=34.5%. The odds ratios show that the flights with the flight distance between 750 to 1000 miles will have higher odds of high delay than other categories.

The odds ratio for the precipitation shows that for each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0392*10)-1= 48.0%.

6.2.2 Classification of delay based on Decision Tree

Both Entropy method and likelihood ratio chi-square were used as measures of split criteria to fit tree models. The results showed that the Entropy method is better according to the misclassification rate and RSC for testing association between the branches and the target categories. As shown in Table 6.6, the maximal tree was pruned back to yield the sequence of terminal nodes. Misclassification rates based on training data are decreasing monotonically as the number of modes increase. However, the misclassification rates based on the validation data show to reach a minimum value for the tree having 5 nodes. Figure 6.1 illustrated the procedure to select the best size tree model. Further, Figure 6.2 shows the tree diagram with 5 terminal nodes.

Table 6.6 Missing Classification Rate and Leaves of Tree Sequence using

Leaves	Training	Validation
1	0.2059	0.2015
2	0.2059	0.2015
3	0.2002	0.1953
4	0.2002	0.1953
**5	0.1977	0.1949**
6	0.1977	0.1949

**Minimum cost tree

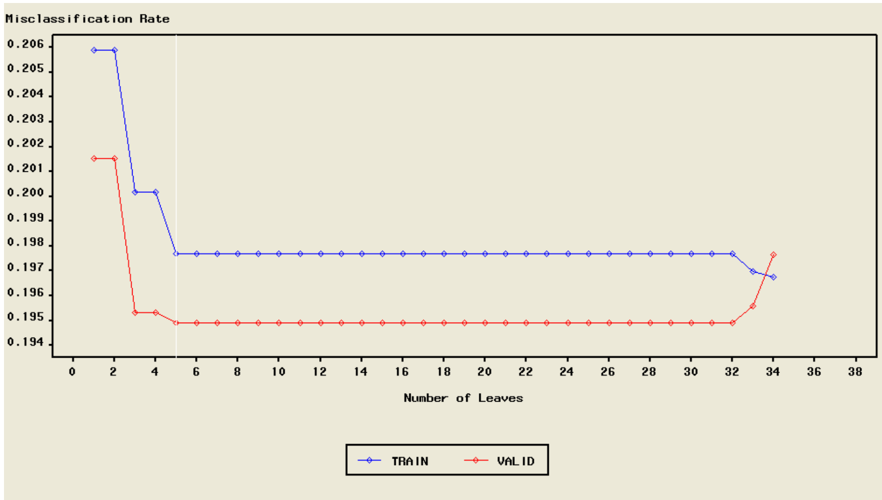

Figure 6.1 The best size tree model based on missing Classification

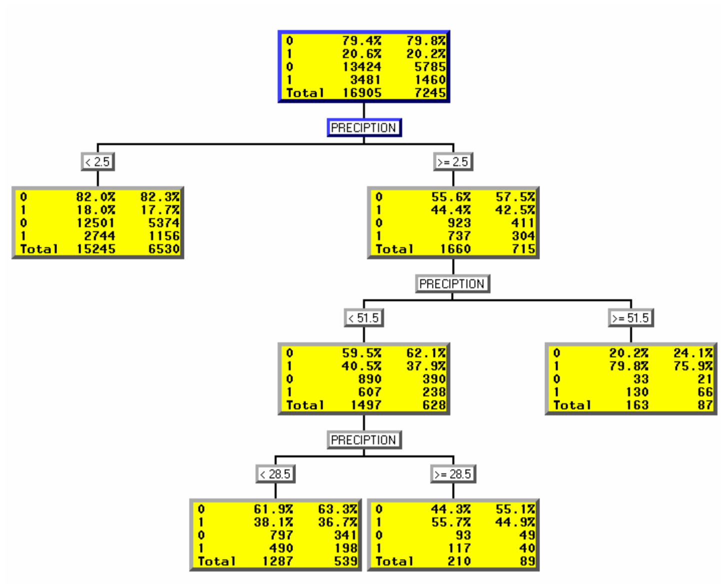

Figure 6.2: Tree classification diagram

According to the decreasing order of variable importance, the tree model shows that the most important variables associated high delay are Precipitation and wind. The corresponding importance values for the variables are shown in Table 6.7.

Table 6.7 Important variable selection base on tree model

Name	Importance	Role
Precipitation	1	input
Wind	0.1682	input

The special contribution of the tree model to the delay analysis is that the complex tree classification is helpful to find the complex pattern based on combined variables. The Table 6.7 illustrates 2 most significant variables, which gives us the information that at present the weather conditions are most important variables for the delay problem due to destination airport at MCO.

6.2.3 Classification of delay based on Neural network

The models in this section are formulated based on data same as before. The data is partitioned into two parts, 70% for training model and 30% for validation.

4 MLP and 3 NRBF neural network models having a range (2 to 5) of hidden nodes are compared. The result from Table 7, the assessment of neural network model, shows that the MLP network with 3 hidden nodes performs best among the models in their respective architectures. This model is compared with the results of logistic regression and tree models. In Figure 6.3, the average error plot for MLP model with 3 hidden nodes shows that when the iteration number increases above 10, the model derived from the training data shows a good fit on the valid data.

Table 6.8 Assessment of neural network model

Neural network	# of nodes	Root ASE	Valid: Root ASE	Misclassification Rate	Valid: Misclassification Rate
MLP	2	0.3923	0.3915	0.1998	0.1964
MLP	3	0.3925	0.3912	0.1995	0.1957
MLP	4	0.3940	0.3915	0.2003	0.1961
MLP	5	0.3925	0.3912	0.1995	0.1957
RBF	3	0.4044	0.4012	0.2059	0.2015
RBF	4	0.4044	0.4012	0.2059	0.2015
RBF	5	0.4044	0.4012	0.2059	0.2015

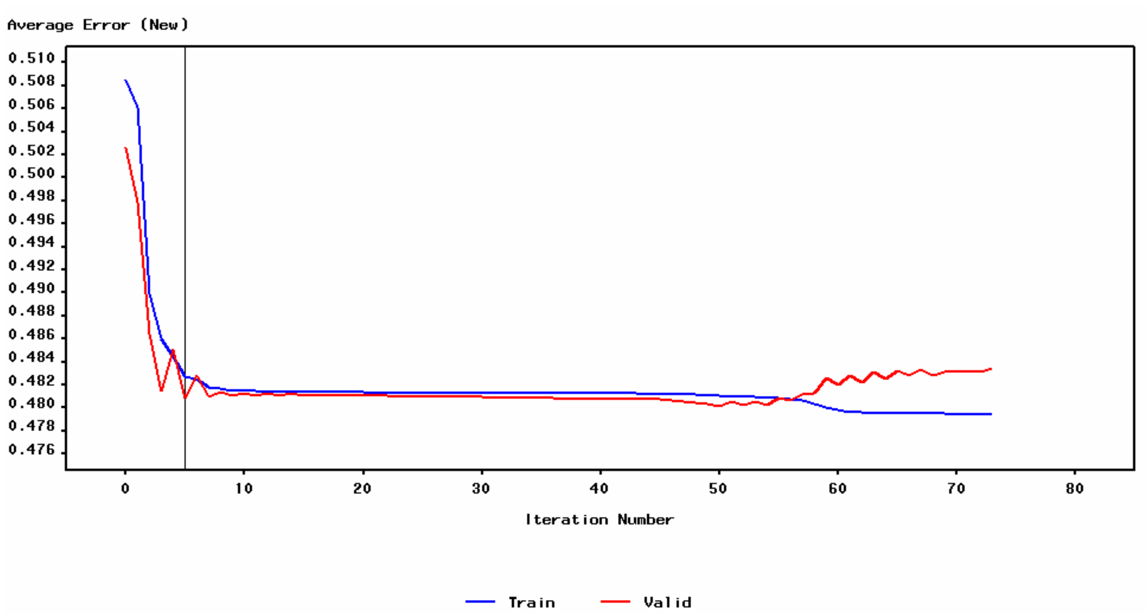

Figure 6.3 Average error plot for MLP model with 3 hidden nodes

6.2.4 Model assessment

In this study, the main purpose of fitting logistic regression model, tree model and neural network model is to predict probability of highly delayed flight occurrence. The correct predicted rate is used to assess model performance. From the Cumulative percentage captured Response Chart as shown in Figure 6.4, MLP neural network model with 3 nodes and logistic regression model performs better than tree model in data prediction.

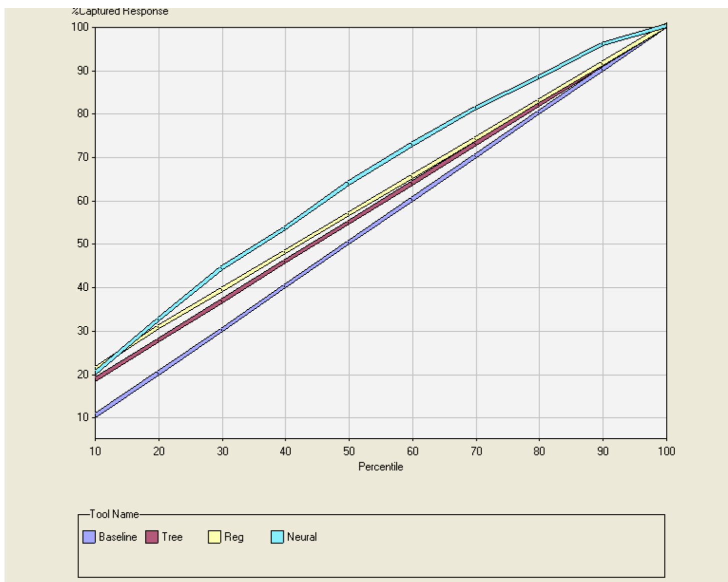

Figure 6.4: Assess model performance: captured response lift plots for 3 models

The captured response lift plots for models of the flight delay are showed in Figure 6.4 above. The best models were identified through the lift plot having cumulative percentage of captured response in the validation dataset on vertical axis. The higher a curve from the baseline curve the better is the performance of the corresponding model.

The performance of each model may be measured by determining how well the models capture the target event across various deciles. From a practical application point of view it must be understood that high delay flights are not normal events and one would need to be parsimonious in issuing warnings for high delay. Therefore, it might be not be reasonable to assign more than 20-30% of observations as high delay and it was decided to evaluate the model performances based on percentage of high delay identified within first three deciles (deciles = 10 percentiles) of posterior probability. It should be noted it (the posterior probability) is not the probability of high delay occurrence at a given point in time but is a measure providing the relative likelihood of high delay occurrence given the composition of the sample. That is the reason in this research we have examined the performance of the models on validation dataset based on percentiles rather than setting a specific threshold on posterior probability. It may be noted that the logistic regression model and MLP model have its captured response percentage higher than the tree model within first 3 deciles (deciles = 10 percentiles).

The result from Table 6.9, the assessment of 3 models, shows that the Root ASE and misclassification rate for each model. The MLP network with 3 hidden nodes and the Logistic regression model are found to perform almost the same as tree model. From the logistic regression model, the variables of arrival time, flight distance, season, region, preciption and wind are found related to the arrival flight. From the tree model, we found that the weather conditions play the most important roles in the flight arrival delay.

Considering the neural network's performance is black box, we will use the logistic regression to analyze the results of prediction.

Table 6.9 Assessment of 3 models

Model	Root ASE	Valid: Root ASE	Misclassification Rate	Valid: Misclassification Rate
Tree model	0.3959	0.3928	0.2002	0.1953
Logistic regression	0.3962	0.3935	0.1996	0.1954
MLP	0.3929	0.3912	0.1995	0.1957

6.3 Conclusions and Discussions

Using the 2002-2003 Airline On-Time Performance Data from Federal Aviation Administration (FAA) and weather data from the National Climatic Data Center (NCDC), this study examined the delay of arriving flights due to destination airport at MCO based on neural network model, logistic regression model and decision tree model.

The models examined the delay pattern related to arrival demand, airport capacity, weather conditions in Orlando, season influences, time influences, week influences, distance influences, regional influences. Seven factors including season influences, time influences, week influences, distance influences, regional influences, wind speed and precipitation show significant association with the likelihood of flights of high delay.

From the results, the highly delay flights show significant seasonal pattern. In summer, the odds of high delay is much higher than in other seasons, while in fall the odds of high delay is much lower than in other seasons. The highly delay flights also show significant daily pattern. During the time from 9pm to 11pm, the flights have the fewer odds to have high delay, while during the time from 3pm to 9pm the flights have the higher odds to have high delay.

Adjusting for other variables, the odds of having higher arrival delay will be about 45-50% higher than other areas when the flight takes off from the east area. For the flights with the flight distance between 750 to 1000 miles, the odds of having a higher arrival delay will be higher than other categories, while for the flights with the flight distance less than 750 miles, the odds of having a higher arrival delay will be the lowest. The variables of precipitation and the wind speed also contribute to the higher delay for arriving flights.

Compared with the results from Chapter 5, the arrival delays due to MCO show apparently regional pattern, besides the seasonal, weekly and daily patterns. And the short-distance flights (<750 miles) show significantly less probability to be delayed more than 15 minutes than mid- and long-distance flights.

The delay propensity analyses in this study provide a better understanding of flight delay problem and provide more information to seek effective countermeasures.

CHAPTER 7 CONCLUSIONS AND DISCUSSION

7.1 General

In this research, statistical models for airport delay and single flight arrival delay are developed. The models use the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the Surface Airways Weather Data from the National Climatic Data Center (NCDC). Multivariate regression, ANOVA, neural networks and logistic regression are used to detect the pattern of airport delay, aircraft arrival delay and schedule performance. These models are then integrated in the form of a system for aircraft delay analysis and airport delay assessment. In this chapter we summarize conclusions from this study. The contributions of this research are also discussed along with the future scope.

7.2 Summary and Conclusions

One of the concerns of this thesis is the delay problem in the context of airports. The delay distribution of an airport can make it easier to understand the airport delay. The assessment of an airport's schedule performance is also discussed. Finally, Multivariate regression, ANOVA, Neural networks and Logistic regression are used to detect the pattern of airport arrival delay.

The results of the research show that the arrival delay is highly related to the originate delay. The airport arrival delay is found to show seasonal and weekly patterns, which is related to the schedule performance. The precipitation and wind speed are also found to be contributors of airport arrival delay. The capacity of airport is not found to be significant. This may indicate that the capacity constraint is not an important problem at MCO.

At the same time this research enables us to investigate the delay at the flight level, and different delay level are compared, which gives out the pattern of arrival delay. Then, the effect of a flight on the immediate flight is considered. We measure the time interval of two consecutive flights and analyze its effect on the flight delay.

The characteristic of single flight and their effect on flight delay are considered. The patterns of delay from the flight level in which delays occur are analyzed, and the significant reasons of delay are given out. The precipitation, flight distance, season, weekday, arrival time and the space between two successive arriving flights are found to contribute to arrival delay of flights. We measure the time interval of two consecutive flights and analyze its effect on the flight delay. The results show that as the spacing between two successive arriving flights increases, the probability of the flights being delayed will decrease.

The characteristics of air congestion that we illustrate empirically is that a flight delay is not necessarily during a peak period, but depends on the impacts generated on subsequent flights. In response to single flight delay predictions and reason for these delays that are generated by the model, which can give indications for the appropriate recovery actions to recover/avoid these delays.

While it was possible to calculate the immediate impact of originat delays, it is not possible to calculate their impact on the cumulative delay. If a late originating aircraft has no space in its down line schedule, it will continue to be late. If that aircraft enters a connecting airport, it can pass its lateness on to other aircraft. So in the research we also consider purifying only the arrival delay at MCO, excluding the flights with originating delay >0. The model makes it possible to see the pattern of the aircraft arrival delay. The weather conditions are found to be the most significant factors that influence the arrival delay due to the destination airport.

7.3 Comments and future research

Delays may also be attributed to airline operations procedures. This type of operation is desirable from an airline point of view because it allows the passengers, aircraft and crew to be rerouted to various destinations. They also provide airlines the opportunity to consolidate passengers into some flights while canceling others. Another factor is the aircraft size. For example, the turboprops require a smaller runway, climbe more slowly and fly at lower altitudes than the jets. These characteristics allow them to be naturally separated from the higher altitude jet traffic. Increased numbers of smaller jets, which operate in the same flight regime as the larger jets, means more aircraft competing for the same airspace, thereby increasing congestion and delays. The airline information and aircraft model are not considered in the thesis make the drawbacks. What is important, due to the lack of data, the airport condition and weather information at the origin airports are not available, which decreases the reliability of the models.

It is theorized that larger airlines should have increased exposure to delays from weather when they serve more destinations. More destinations mean more potential delays to be spread throughout the system. Because they have more flights, larger airlines can more accurately predict the likelihood of crew sickness or mechanical failure. This enables them to keep a smaller percentage of their resources in reserve than smaller carriers, while maintaining the same on-time performance. Delay data on both large- and medium-sized national carriers allows comparison of the effects of congestion and weather delays on airlines with different network characteristics (Rosen 2002). These can be a direction of future research.

APPENDIX A: VARIABLES USED IN THE REGRESSION MODELS

Variables	Definition
Flight arrival delay	Difference of the actual arrival time minus the scheduled arrival time
Maximum hourly flow rate	Maximum numbers of operation that can be handled in a one-hour period under specific operating conditions
Arrival demand	Number of completed arrival flights to MCO per day according to the scheduled arrival time.
Flight duration	Airborne time for each flight
Space of Inter-arrival time	Intervals between two consecutive arriving flights
Airport precipitation	Daily observations about the inches of the rainfall
Airport wind speed	Daily average wind speed at MCO, speed of wind in mph per day.
Seasonal variables	Indicate the seasons when the flights are scheduled, spring (March-May), summer (June-August), fall (September to November), and winter (December to February).
Weekly variables	Indicate the weekday when the flights are scheduled
Time variables	Indicate the scheduled arriving time of each delayed flights, morning (7am to 11:59 am), afternoon (12 am to 4:59 pm), and evening (5 pm to 11:59 pm).
Origin airport regional variables	The regional effects are captured by a set of dummy variables, south, east, central, and west areas (definition is in Appendix B).
Flight distance	Categories of flight distance, which respectively represent the distance of 0 to 750 miles, 750 to 1000miles and greater than 1000 miles.

APPENDIX B: DEFINITION OF REGIONAL VARIABLES

Areas	Definition
South	States of Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee
East	States of Connecticut, Delaware, District of Columbia, Indiana, Kentucky, Maine, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, New York, Ohio, Pennsylvania, Rhode Island, Vermont, Virginia, and West Virginia
Central	States of Arkansas, Colorado, Illinois, Iowa, Kansas, Louisiana, Minnesota, Missouri, Nebraska, North Dakota, Oklahoma, South Dakota, Texas, and Wisconsin
West	States of Alaska, Arizona, California, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming

REFERENCES

1 Abdelghany, K. F., Abdelghany, A. F., and Raina S., (2004) A model for projecting flight delays during irregular operation conditions, Journal of Air Transport Management, Volume 10, Issue 6, Pages 385-394

2 Aisling, R., and J.B. Kenneth, (1999) An assessment of the capacity and congestion levels at European airports, ERSA conference papers ersa 99, pages 241, European Regional Science Association.

3 Allan, S.S., S.G. Gaddy, and J.E. Evans, (2001) Delay Causality and Reduction at the New York City Airports Using Terminal Weather Information, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Lexington, Massachusetts

4 Allison, P. D., (1999) Logistic Regression using the SAS System, John Wiley & Sons, Inc,

5 Ashford and Wright, (1992) Airport Engineering, John Wiley & Sons, Inc,

6 Bureau of Transportation Statistics, Airline On-Time Statistic. U.S. Department of Transportation. Washington, D.C. http://www.bts.gov/programs/airline_information.

7 Bracciali, C. F., X. Li, and A. S. Ranga, (2005) Real orthogonal polynomials in frequency analysis, Math. Comp. Volume 74, pages 341-362.

8 Christodoulou, C., and Georgiopoulos, M., Applications of Neural Networks in Electromagnetics, Artech House, Boston, 2001.

9 Daruis, L., O. Njåstad and W. Van Assche, Para-orthogonal polynomials in frequency

analysis, Rocky Mountain J. Math., volume 33, pages 629-645.

10 Girden, E. R., ANOVA: repeated measures, Newbury Park, Calif., Sage Publications, 1992.

11 Hagan, M., T., and Menhaj, M., (1994) Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks, Volume 5, No. 6, pages 989-993.

12 Hansen, M., and C. Y. Hsiao (2005), Going South? An Econometric Analysis of US Airline Flight Delays from 2000 to 2004, Presented at the 84rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2005.

13 Hansen, M., and D. Peterman, (2004) Throughput Impacts of Time-based Metering at Los Angeles International Airport, Presented at the 83rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2004.

14 Hansen, M., S. J. Tsao, A. Huang, and W. Wei, Empirical Analysis of Airport Capacity Enhancement Impacts: A Case Study of DFW Airport, presented at the 1999 Transportation Research Board Annual Meeting, Washington, D.C., 1999.

15 Hansen, M., (2002) Micro-level analysis of airport delay externalities using deterministic queuing models: a case study, Journal of Air Transport Management Volume 8, Issue 2 , Pages 73-87.

16 Hansen, M., Y. Zhang, Operational Consequences of Alternative Airport Demand Management Policies: The Case of LaGuardia Airport, Presented at the 84rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2005.

17 Hosner, D., W., and S. Lemeshow, Applied Logistic Regression, Wiley & Sons, 1989.

18 http://www.travelforecast.com

19 Janic, M., Large-schale Disruption of an Airline Network: a Model for Assessment of the Economic Consequences, Presented at the 82rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2003.

20 Jones, W.B., O. Njåstad and E.B. Saff, (1990) Szego polynomials associated with Wiener-Levinson filters, J. Comput. Appl. Math., volume 32, pages 387-407.

21 Kanafani, A., M. R. Ohsfeldt, and W. J. Dunlay, (2004) Comprehensive Evaluation of Investments – A new method of evaluating impacts of technologies in air traffic management, Presented at the 83rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2004.

22 Mehndiratta, S.R., M. Kiefer, and G. C. Eads, Analyzing the Impact of Slot Controls: The Case of San Francisco International Airport, Presented at the 81rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2002.

23 Mueller, E.R. and G. B. Chatterji, (2002) Analysis of Aircraft Arrival and Departure Delay Characteristics, AIAA's Aircraft Technology, Integration, and Operations (ATIO), Los Angeles, California

24 Neural Networks, http://www.statsoft.com/textbook/stathome.html

25 Pande, A., (2005) Applying Hybrid Models for Real-time Crash Risk Assessment of Freeways, Department of Civil and Environmental Engineering, University of Central Florida PHD dissertation

26 Rosen, A. (2002), Flight Delays on US Airlines: The Impact of Congestion Externalities in Hub and Spoke Networks, Department of Economics, Stanford University

27 Schaefer, L., and D. Miller, (2001) Flight Delay Propagation Analysis with the Detailed Policy Assessment Tool, Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference

28 Suzuki, Y., (2000), The relationship between on-time performance and airline market share: a new approach, Transportation Research Part E, volume 36, pages 139-154.

29 The introduction to SAS Enterprise Miner Software, available from SAS.

30 Wang, P. T., C. R. Wanke, F. P. Wieland, (2004) Modeling Time and Space Metering of Flights in the National Airspace System, Proceedings of the 2004 Winter Simulation Conference, Washington, DC.

31 Wu, C. (2005), Inherent delays and operational reliability of airline schedules, Journal of Air Transport Management Volume 11, Issue 4 , Pages 273-282

32 Yan, S., C.Y. Shieh and M. Chen, (2002) A simulation framework for evaluating airport gate assignments, Transportation Research Part A: Policy and Practice Volume 36, Issue 10 , Pages 885-898.