University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Computer Science	Spring semester, 2014 Open / Restricted access
Writer: Jie Cheng	………………………………………… (Writer's signature)
Faculty supervisor: Prof. Chunming Rong	
Thesis title: **Risk Management Using Big Real Time Data**	
Credits (ECTS): 30	
Key words: flight delay prediction, smoothing spline, ARIMA, multiple regression, weather effect, Java, R analysis, maven, web application	Pages: 77 + enclosure: Stavanger, 15 June, 2014 Date/year

Front page for master thesis
Faculty of Science and Technology
Decision made by the Dean October 30th 2009

UNIVERSITY OF STAVANGER

MASTER THESIS

Risk Management Using Big Real Time Data

Author:
Jie Cheng

Supervisor:
Prof. Chunming Rong

A thesis submitted in fulfillment of the requirements
for the degree of Master in Computer Science

June 15, 2014

Acknowledgements

I would like to thank my supervisor, Prof. Chunming Rong for his valuable and interesting suggestions during the process of preparing this thesis. Besides, I also appreciate the necessary facilities he has given me to accomplish my Master's thesis.

At the same time, I would like to express my deep gratitude to Prof. Zhou and Prof. Chen who were academic visitors in Uis. They always assisted me when I have confusions. Meanwhile, also thank Antorweep, PhD student at UiS, for his nice and abundant help to me.

At last, I would like to thank my family, friends and specially my husband for his unconditional support no matter on study or other decisions.

Contents

List of Tables

List of Figures

UNIVERSITY OF STAVANGER

Abstract

Department of Electrical Engineering and Computer Science

Master of Computer Science

Risk Management Using Big Real Time Data

by Jie Cheng

Adding to societal changes today, are the miscellaneous big data produced in different fields. Coupled with these data is the appearance of risk management. Admittedly, to predict future trend by using these data is conducive to make everything more efficient and easy. Now, no matter companies or individuals, they increasingly focus on identifying risks and managing them before risks. Effective risk management will lead them to deal with potential problems.

This thesis focuses on risk management of flight delay area using big real time data. It proposes two different prediction models, one is called General Long Term Departure Prediction Model and the other is named as Improved Real Time Arrival Prediction Model. By studying the main factors lead to flight delay, this thesis takes weather, carrier, National Aviation System, security and previous late aircraft as analysis factors. By utilizing our models can do not only long time but also short term flight delay predictions. The results demonstrate goodness of fit. Besides the theory part, it also presents a practical and beautiful web application for real time flight arrival prediction based on our second model.

CHAPTER 1 INTRODUCTION

1.1 Background and Motivation

The succession of rapid data increases and computational ability lead to a fast development of data mining. Competitive companies or research institutions collect massive volume of data (usually called Big Data) to do data analytics.

Effective data mining algorithms and analysis strategies can extract precious information for companies or individuals to gain pre-knowledge to make a further decision. Among those fields with big data, one of them has aroused extensively attention, which is flights delay predictions. Great importance of risk management of flight delays can be seen in recent years. The appalling MH370 flight accident happened this year pushes flight risk management to an extremely urgent situation. Besides, 19% of the US domestic flights delayed more than 15 minutes. Tremendous economy cost and dissatisfaction have been brought to airline companies and passengers. So no matter from the safety factor or the economy side, more effective flight delay prediction models should be developed and improved.

In order to establish a suitable prediction model, this thesis explored and compared miscellaneous mathematic methods. After studying those methods, this thesis aims to build novel model for the predictions of flight delays using big real time data like weather, carriers, airports and also large historical data. Furthermore, the second model will be implemented through a website where users can explore the model and check the status of a specific flight.

1.2 Related Work

A lot of researches have also been conducted on the management and propagation of flight Normal or Poisson distributions, which aims at improving traffic management systems. Mueller and Chatterji [1] just made a model based on Normal or Poisson distributions to simulate departure, en-route and arrival delays. But those models are too general to concern about flights or airlines features. Zonglei et al [2] demonstrates predictions of percentage of delayed flights on an airport using decision trees and neural networks.

Besides, in recent years, Bayesian Network (BN) models have been proposed with different improved algorithms, based on parameter learning, structure learning, and some mixed algorithms [3] [4]. BN is a machine learning method based on graph and probability theory, which is an efficient method for modeling and estimating complicated situations [5]. The benefit of Bayesian theory is it not only based on historical data but also priori probability. However, there is a lack of priori probability for the delay of a flight model.

Tu & Ball [6] applied general spline function and a modified genetic algorithm for estimating the departure delay distribution. The model consists of a seasonal trend, a daily trend and a random residual. The whole system is complex and seems expensive to compute especially for the residual part with genetic method. And they only generate a general arrival delay model for all flights regardless of current weather effect.

Based on Tu& Ball's work, Vincent Martinez, who is a master student and specifically focused on customer long-term information by using kernel density estimation method. However, even the optimal models with the most relevant parameters have been selected to implement predictions with large amount of data; it still has not considered real-time factors like weather influence. Some severe weather conditions will be the determinant in some situations.

In addition to the academic research area of prediction models, some mobile applications or websites also started to provide flight status check services. For example, website FlightCaster provide probabilities of a flight being on-time, less than one hour late or more than one hour late by utilizing airports, airlines, weather and historical data. Nevertheless, their model doesn't predict the estimated arrival delay minutes instead of a general delay probability on the three defined delay intervals.

1.3 Contribution

Compared to those research models and website applications of flight delay predictions, this project mainly focus on more reasonable, economical but novel models, especially our second model which is using big real time data to implement flight risk management. Specifically, it has the following characteristics:

- Using latest big real time weather data for each flight instead of a global trend.

  For real time weather data, this thesis not only uses basic weather indicators as temperature, precipitation, etc. Instead, it utilizes three major weather factors---wind speed, visibility and sky conditions, which will be introduced with details in following chapters.

- Establishing a high-efficiency and low running time model.

  This thesis utilized a Smoothing Spline function combined with a multiple linear regression model to do data trainings and predictions.

- Analyzing one specific airport and airline to clearly show how the project puts theory into practice.

  In this thesis, we explore San Francisco International Airport in United States and American Airline specifically. All data are downloaded from The Bureau of Transportation Statistics.

- Implementing a user-friendly web site to make flight search possible by every user.

  This website currently provides flight arrival delay prediction of American Airline in San Francisco International Airport which is parallel with the data we have trained. It is also possible to explore all airlines at all airports, which just need to update and extend the database. The aim of showing one airline in one airport here is just to show the methods the second model has used. Comprehensive search functions can be developed in future works.

1.4 Thesis Structure

Chapter 2.	Data Exploration
Chapter 3.	Mathematic Theory Background
Chapter 4.	Model Description and Implementation
Chapter 5.	Data Application & Testing
Chapter 6.	Model Evaluation
Chapter 7.	Web Application Implementation & Performances
Chapter 8.	Conclusions & Future Work

CHAPTER 2 DATA EXPLORATION

Before introducing our prediction models in this thesis, we will explain the dataset has been used. This aims to give readers a better idea to understand our model and the principles we have used according to the dataset characteristics. From the dataset structure, we can see delay factors apparently.

2.1Dataset Overview

The dataset we have used in this thesis is from the Bureau of Transportation Statistics (BTS), which is a publicly accurate data collected by all USA domestic flights of major airlines:

AirTran Airways (FL)
Alaska Airlines (AS)
American Airlines (AA)
American Eagle (MQ)
Delta Air Lines (DL)
ExpressJet Airlines (EV)
Frontier Airlines (F9)
Hawaiian Airlines (HA)
JetBlue Airways (B6)
SkyWest Airlines (OO)
Southwest Airlines (WN)
United Airlines (UA)
US Airways (US)
Virgin America (VX)

Each airline reports monthly to the BTS about numbers of flight delays and also began reporting causes of delays in June 2003. Each flight data are composed of the following information:

- Flight Information:

Carrier, Flight Number, Tail Number

- Time Information:
  Date, scheduled departure/arrival time, actual departure/arrival time, scheduled duration, actual duration, departure delay and arrival delay

- Airport Information:
  Departure airport, arrival airport, taxing duration, take-off and landing hour

- Delay Reason Information:

Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.

Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control such as crew problems or maintenance, cleaning, baggage loading or fueling, etc.

National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.

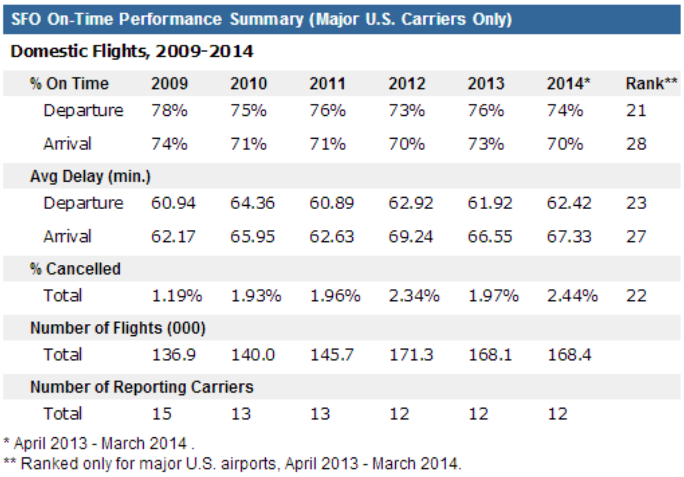
Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

2.2Dataset Characteristics

In order to build a reasonable and economical model based on datasets published by the BTS, it is indispensible to study characteristics of these data. After some tests conducted, some significant features have been showing as below.

●Airport Pattern

Airports differ from each other in number of flights, average delay time, carrier numbers and cancellation rate. These disparities can be shown apparently from the following Table 2.2A [7] and Table 2.2B [8] according to official data from the BTS. Here we just take San Francisco International Airport and Denver International Airport as examples.



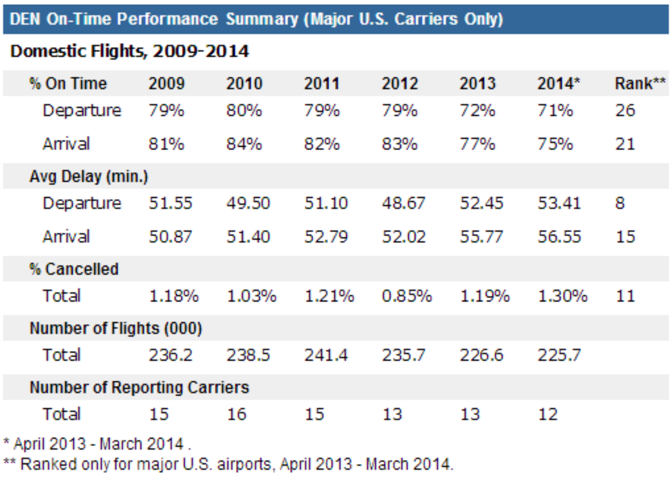
SFO On-Time Performance Summary (Major U.S. Carriers Only)

Domestic Flights, 2009-2014

% On Time	2009	2010	2011	2012	2013	2014*	Rank**
Departure	78%	75%	76%	73%	76%	74%	21
Arrival	74%	71%	71%	70%	73%	70%	28
Avg Delay (min.)							
Departure	60.94	64.36	60.89	62.92	61.92	62.42	23
Arrival	62.17	65.95	62.63	69.24	66.55	67.33	27
% Cancelled							
Total	1.19%	1.93%	1.96%	2.34%	1.97%	2.44%	22
Number of Flights (000)							
Total	136.9	140.0	145.7	171.3	168.1	168.4	
Number of Reporting Carriers							
Total	15	13	13	12	12	12	

* April 2013 - March 2014 .
** Ranked only for major U.S. airports, April 2013 - March 2014.

TABLE 2.2A: SAN FRANCISCO INTERNATIONAL AIRPORT



DEN On-Time Performance Summary (Major U.S. Carriers Only)

Domestic Flights, 2009-2014

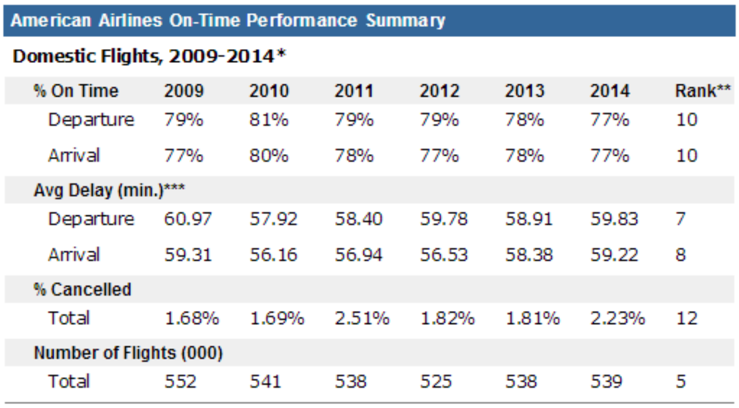
% On Time	2009	2010	2011	2012	2013	2014*	Rank**
Departure	79%	80%	79%	79%	72%	71%	26
Arrival	81%	84%	82%	83%	77%	75%	21
Avg Delay (min.)							
Departure	51.55	49.50	51.10	48.67	52.45	53.41	8
Arrival	50.87	51.40	52.79	52.02	55.77	56.55	15
% Cancelled							
Total	1.18%	1.03%	1.21%	0.85%	1.19%	1.30%	11
Number of Flights (000)							
Total	236.2	238.5	241.4	235.7	226.6	225.7	
Number of Reporting Carriers							
Total	15	16	15	13	13	12	

* April 2013 - March 2014 .
** Ranked only for major U.S. airports, April 2013 - March 2014.

TABLE 2.2B: DENVER INTERNATIONAL AIRPORT

●Airline Pattern

Similar features as airport pattern have been found as following Table 2.2C [9] and Table 2.2D [10]. We can easily see American Airlines have a much higher departure/arrival delay rate and longer delay time than Southwest Airlines.

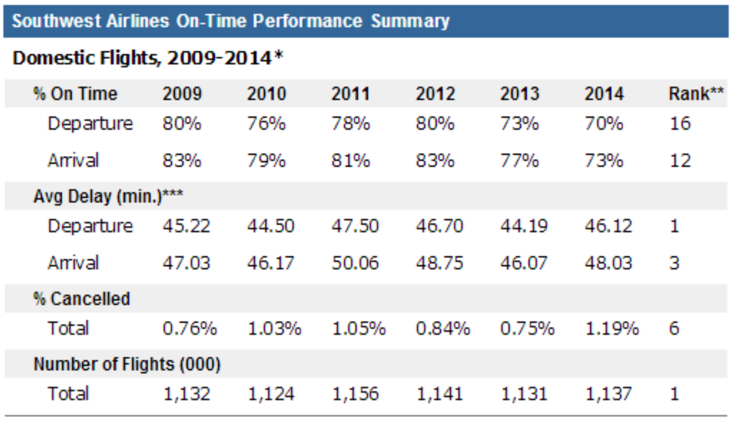


American Airlines On-Time Performance Summary

Domestic Flights, 2009-2014*

% On Time	2009	2010	2011	2012	2013	2014	Rank**
Departure	79%	81%	79%	79%	78%	77%	10
Arrival	77%	80%	78%	77%	78%	77%	10
Avg Delay (min.)***							
Departure	60.97	57.92	58.40	59.78	58.91	59.83	7
Arrival	59.31	56.16	56.94	56.53	58.38	59.22	8
% Cancelled							
Total	1.68%	1.69%	2.51%	1.82%	1.81%	2.23%	12
Number of Flights (000)							
Total	552	541	538	525	538	539	5

TABLE 2.2C: AMERICAN AIRLINES



Southwest Airlines On-Time Performance Summary

Domestic Flights, 2009-2014*

% On Time	2009	2010	2011	2012	2013	2014	Rank**
Departure	80%	76%	78%	80%	73%	70%	16
Arrival	83%	79%	81%	83%	77%	73%	12
Avg Delay (min.)* **							
Departure	45.22	44.50	47.50	46.70	44.19	46.12	1
Arrival	47.03	46.17	50.06	48.75	46.07	48.03	3
% Cancelled							
Total	0.76%	1.03%	1.05%	0.84%	0.75%	1.19%	6
Number of Flights (000)							
Total	1,132	1,124	1,156	1,141	1,131	1,137	1

TABLE 2.2D: SOUTHWEST AIRLINES

- Weather Influence

Firstly, Figure 2.2A [11] shows the percent of five factors of total delay based on classifications and statistic analyses of the BTS. Here, we only can see Extreme Weather factor has an average 4 percent of flight delays and Aircraft Arriving Late becomes the most vital reason of flight delays.

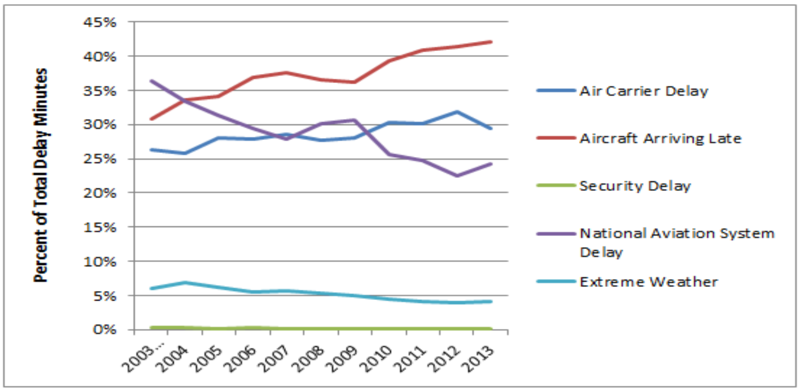


FIGURE 2.2A: PERCENT OF TOTAL DELAY MINUTES

But according to explanations of the BTS, 4 percent is only caused by extreme weather conditions. Less extreme weather conditions can also cause flight delays. So now, we wonder what is the actual weather's share of total delay?

According to the BTS, there is another category of weather within the NAS category. This factor doesn’t prevent flying but slows down the operations of the system. Figure 3.2B [12] shows the real weather’s share of total flight delays.

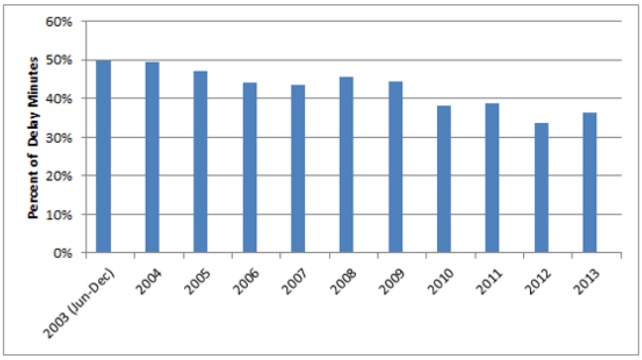


FIGURE 3.2B: WEATHER’S SHARE OF TOTAL DELAY MINUTES

From the above figure, it is apparently that weather has a large share of total delays, which means it will has significant influence if we can figure out how weather delays flights.

CHAPTER 3 MATHEMATIC THEORY BACKGROUND

3.1Smoothing Spline Estimation

3.1.1 Definition of Spline

A spline [13] is a piece-wise polynomial with pieces defined by a sequence of knots

$\xi_1 < \xi_2 < \xi_{3 \ldots} < \xi_k$

such that the pieces join smoothly at the knots.

A spline of degree m can be represented as a power series:

$$S(d) = \sum_{j=0}^{m} \beta_j D^j + \sum_{j=1}^{k} \lambda_j (d - \xi_k)^m{}_+$$

where the notation

$$(d-\xi_k)_+ = \begin{cases} d - \xi_k, d > \xi_k \\ 0, otherwise \end{cases}$$

3.1.2 Smoothing Splines

It is a method to consider fitting a spline with knots at every data point (same weight), which means it could potentially fit perfectly with estimation of its parameters by minimizing the usual sum of squares plus a roughness penalty.

Usually a suitable penalty is to integrate the squared second derivative, known as the following formula:

$$PSS = \sum_{d=1}^{n} (y_d - S(d))^2 + \lambda \int (S''(d))^2 dx$$

where λ is a tuning parameter. y_d denotes the average day delay and can be calculated by

$$y_d = \frac{\sum_t \sum_i y_{dti}}{\sum_t n_{dt}} \qquad d=1, 2, 3, \ldots, n$$

where y_{dti} denotes the observed delay of flight i, at time t and on day d; n_{dt} denotes number of flights at time t on day d. For PSS function, if λ → 0, it means there is no penalty and the spline can have a very close fit, but the result curve could be very noisy as it follows every detail in the data. If λ →∞, the effect of the spline is the opposite. The spline can be very smoothly, but may be a bad fit.

3.2ARIMA(p,d,q) Model

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model [14]. An ARMA model consists of Moving averages (MA) and autoregressive (AR) processes. The difference between ARMA and ARIMA is ARMA is used for forecasting stationary time series and ARIMA is designed for non-stationary time series. ARIMA is the combination of difference operation with ARMA.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:
p is the number of autoregressive terms,
d is the number of nonseasonal differences, and
q is the number of lagged forecast errors in the prediction equation.

It follows the formula below:

$$\begin{cases} \Phi(L)\Delta^d y_t = c + \Theta(L)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2, E(\varepsilon_t, \varepsilon_s) = 0, s \neq t(1) \\ Ey_s\varepsilon_t = 0, \forall\, s < t \end{cases}$$

where $\Delta^d = (1 - L)^d$,

$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^P$, which is the auto regression (AR) coefficient polynomials;

$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$, which is the moving average (MA) coefficients polynomials;

$\{\varepsilon_t\}$ is zero mean white noise sequence

3.2.1 ARIMA Modeling Steps

Time series modeling is based on the hypothesis that the random sequence is stationary. Hence, stationarity of a time series is a vital precondition of modeling. Both ARMA and ARIMA are built on stationary time series. And any non-stationary time series can be transformed to stationary series by suitable sequence operations. After the series turning stationary, we could apply ARMA to model. Details of ARIMA modeling is as follows:

1) Data Stationary Test (ADF)

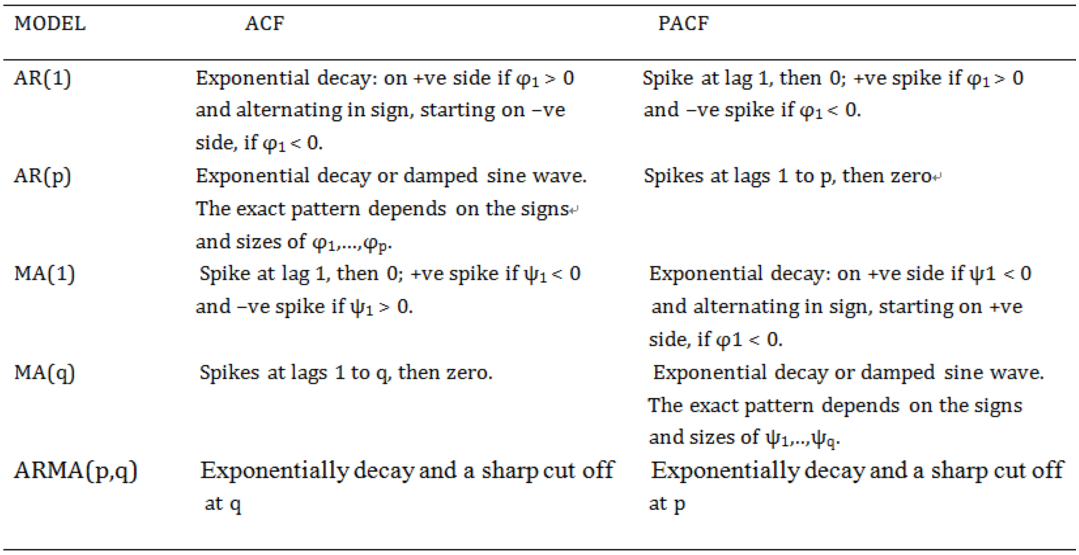
First, we need to do data stationary test. A simple way to see is plotting the time series and evaluating the pattern of the graph. A more accurate method is to apply ADF unit root test.

As to non-stationary time series, if there is an increasing or a decreasing trend, logarithm and difference operations need to be done first. Then ADF test can be applied again. Repeat the above steps until the time series become stationary. The number of difference operations in total is counted as d in ARIMA (p,d,q) model.

2) Fitting Stationary Time Series to ARMA

We use {ys} to denote stationary time series after difference operations. In order to model {ys} with ARMA and find out the parameters p and q, we need to calculate the autocorrelation function(ACF)and the partial autocorrelation function (PACF) of the time series sample. The suitable values of p and q will be chosen from the pattern of ACF and PACF tests. The principles will be described in the following Table 3.1A [15]:

TABLE 3.1A: SHAPES OF ACF AND PACF TO IDENTIFY ARMA MODELS



MODEL	ACF	PACF
AR(1)	Exponential decay: on +ve side if $\varphi_1 > 0$ and alternating in sign, starting on −ve side, if $\varphi_1 < 0$.	Spike at lag 1, then 0; +ve spike if $\varphi_1 > 0$ and −ve spike if $\varphi_1 < 0$.
AR(p)	Exponential decay or damped sine wave. The exact pattern depends on the signs and sizes of $\varphi_1,...,\varphi_p$.	Spikes at lags 1 to p, then zero
MA(1)	Spike at lag 1, then 0; +ve spike if $\psi_1 < 0$ and −ve spike if $\psi_1 > 0$.	Exponential decay: on +ve side if $\psi 1 < 0$ and alternating in sign, starting on +ve side, if $\varphi 1 < 0$.
MA(q)	Spikes at lags 1 to q, then zero.	Exponential decay or damped sine wave. The exact pattern depends on the signs and sizes of $\psi_1,..,\psi_q$.
ARMA(p,q)	Exponentially decay and a sharp cut off at q	Exponentially decay and a sharp cut off at p

3) Selecting Parameters

After testing different combinations of p and q, AIC [16] and SC [17] standards can be applied to choose the best model parameters.

4) Model Verification

Certain diagnostics are used to check the validity of the model. An accurate model should have extracted sufficient information of the time series. One of ways to achieve this goal is to examine whether the residuals are a white noise sequence. If the model cannot pass through this test, which means the residuals are not a white noise sequence, a new model is required again. If the residuals are a white noise sequence, a valid model is found.

5) Model Forecast

According to the model has been selected, a future trend of the time series can be found through suitable statistical software.

3.3Multiple Regression

3.3.1 Related Definitions[18]

The Correlation Coefficient: It indicates whether a relationship exists between two variables, how strong that relationship is and whether these two variables are positively or negatively related.

The Coefficient of Determination: It explains how much variation in one variable is directly related to variation in another variable.

Linear Regression: A process that allows you to make predictions about variable Y based on knowledge you have about variable X.

The Standard Error of Estimate: It presents how accurate your predictions are likely to be when you perform regression analysis.

Multiple Regression: The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. Hence it allows you to examine how multiple independent variables are related to a dependent variable. And we categorize multiple regression into Linear Regression and Nonlinear Regression.

3.3.2 Regression Equation:

1) Linear Regression Equation[19]:

Given a data set $\{\mathcal{y}i, \mathcal{X}i1, \ldots, \mathcal{X}ip\}_{i=1}^{n}$of n stati$i$ts, a linear model assumes the dependent variable y_i is linearly related to X_i. This relationship is reflected through a disturbance term ε_i, which is an unobserved random variable that adds noise to the linear relationship. And the equation has the following form:

$$\mathcal{y}i=\beta_1\ \mathcal{X}i1\ + \ldots +\ \beta_p\ \mathcal{X}ip\ +\varepsilon_i = \boldsymbol{\mathcal{X}i'\ \beta} + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\mathcal{X}i$ is a (row) vector of predictors for the ith of n observations, usually with a 1 in the first position representing the regression constant; β is the vector of regression parameters to be estimated; and εi is a random error, assumed to be normally distributed, independently of the errors for other observations, with expectation 0 and constant variance: $\varepsilon_i \sim \mathrm{NID}(0, \sigma^2)$.

2) Nonlinear Regression Equation[20][21]:

In the more general normal nonlinear regression model, the function f(·) relating the response to the predictors is not necessarily linear:

$$\mathcal{y}i = f(\beta, \mathcal{X}i') + \varepsilon_i$$

As we know, in linear regression, β is a vector of parameters and $\mathcal{X}i'$ is a vector of predictors, but in nonlinear regression, these vectors may not have the same dimension, and ε_i ~ NID(0, σ^2). And the likelihood of nonlinear regression model is as following:

$$\mathcal{L}(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}[\mathcal{y}i - f(\beta, \mathcal{X}i')]^2}{2\sigma^2}\right\}$$

When the sum of squared residuals is minimized, the likelihood is maximized:

$$\mathcal{S}(\beta) = \sum_{\mathrm{i=1}}^{\mathrm{n}}[\mathcal{y}i - f(\beta, \mathcal{X}i')]^2$$

$$\frac{\partial\mathcal{S}(\beta)}{\partial\beta} = -2\sum[\mathcal{y}i - f(\beta, \mathcal{X}i')]\frac{\partial f(\beta, \mathcal{X}i')}{\partial\beta}$$

By setting the partial derivatives to 0, we can get equations for regression coefficients.

3.3.3 Properties of Multiple Regression

In practice we will build the multiple regression model from the sample data using the least squares method. Thus we seek coefficients $\boldsymbol{b_j}$ such that

$$\mathrm{y} = \mathrm{b_0} + \mathrm{b_1}x_1 + \mathrm{b_2}x_2 + \cdots + \mathrm{b_p}x_\mathrm{p}$$

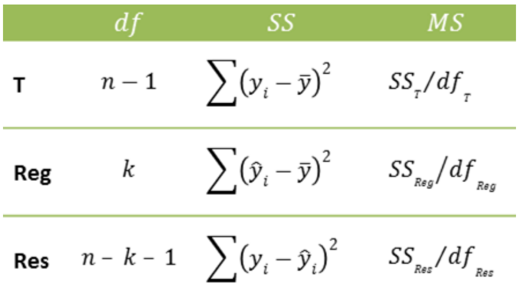
For real data we will have

$$\hat{\mathrm{y}} = \mathrm{b_0} + \mathrm{b_1}x_{i1} + \mathrm{b_2}x_{i2} + \cdots + \mathrm{b_p}x_{i\mathrm{p}}$$

where $\hat{y}_i$ is the y value predicted by the model for the sample data $\boldsymbol{x}_{i1}$, …, $\boldsymbol{x}_{ik}$. Thus the $\boldsymbol{i}$th error term for the model is given by

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip})$$

Depending on the following table and basic mathematic knowledge:



	df	*SS*	*MS*
T	$n-1$	$\sum (y_i - \bar{y})^2$	SS_T/df_T
Reg	k	$\sum (\hat{y}_i - \bar{y})^2$	SS_{Reg}/df_{Reg}
Res	$n - k - 1$	$\sum (y_i - \hat{y}_i)^2$	SS_{Res}/df_{Res}

Property 1:

$$\mathrm{T} = \mathrm{Reg} + \mathrm{Res}$$

$$\mathrm{SS_T} = \mathrm{SS_{Reg}} + \mathrm{SS_{Res}}$$

$$\mathrm{df_T} = \mathrm{df_{Reg}} + \mathrm{df_{Res}}$$

Property 2: Where ***R*** is the multiple correlation coefficient

$$\mathrm{SS_{Res}} = \mathrm{SS_T}(1 - \mathrm{R}^2)$$

$$\mathrm{R}^2 = \mathrm{SS_{Reg}}/\mathrm{SS_T}$$

$$\mathrm{R} = \mathrm{r}_{\mathrm{y}\hat{\mathrm{y}}}$$

$$\mathrm{R}^2 = \frac{\mathrm{SS_{Reg}}}{\mathrm{SS_T}} \leq 1$$

CHAPTER 4 MODEL DESCRIPTION AND IMPLEMENTATION

Previously, we have explained important factors which contribute to flight delays and we also showed useful mathematic fitting functions for flights data. In this chapter, we are going to illustrate two models have been built during the whole projects. The first one is a generally basic **long term departure** prediction model and second one is our final **real time arrival** prediction model for each specific flight, which has been implemented through a web application in our project.

4.1First General Long Term Departure Prediction Model

In this thesis, study is conducted in San Francisco International Airport, which is one of the busiest airports in United States. And as the first model, we focus on building a general model which can predict flight departure delays for all flights regardless of airlines, extreme weather conditions, last late aircraft and etc.

Hence in first basic model, we consider three main categories as the reasons lead to delay. The first category is weather conditions and holiday effect. The second one is time-schedule effect. And the last one is random delay factor effect. (Fig. 4.1 A)

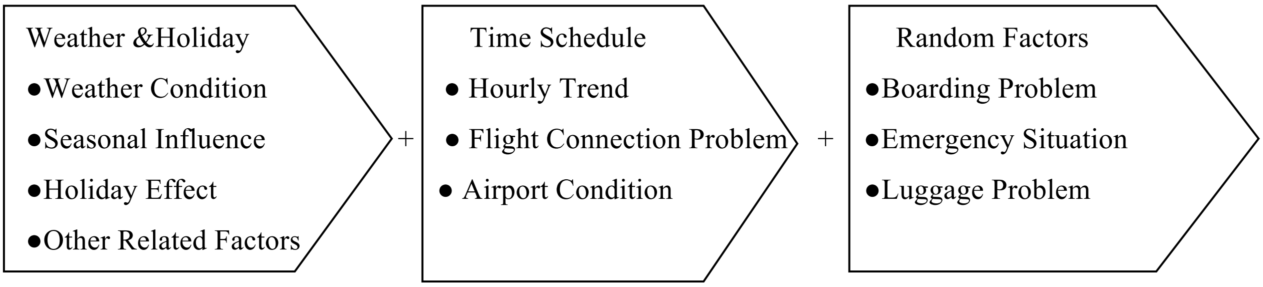


FIGURE 4.1A: GENERAL LONG TERM DEPARTURE PREDICTION MODEL STRUCTURE

By transform these factors to our model; we define the formula as follows:

$$\varphi_{dhtn=}\ \omega(dh) + \delta(t) + \in_n$$

φ_{dhtn} is the departure delay of flight n at schedule time t, on day d and holiday h if the day is. $\omega(dh)$ is the weather and holiday effect; $\delta(t)$ is the daily time-schedule effect and $\in_n$ represents the random effects.

4.1.1 Weather and Holiday Function

In this model, weather effect indicates a general weather trend in different seasons, which doesn't not refer to every day's corresponding weather condition. Here we employ a weighed smoothing spline to fit the weather and holiday pattern. The difference between the weighted spline and the normal spline is we use weight for each knot that connected smoothly to construct the spline.

As stated in Chapter 2, normal smoothing splines will set weights for all knots as 1. But actually in some circumstances, it may not be so accurate to weigh every knot as same weight which means the importance of every knot is different from each other. And from Figure 4.1B Holiday Effect of 2012, we can easily see the delays on holidays or festivals especially important ones(ex.Christmas) are much larger than normal days.

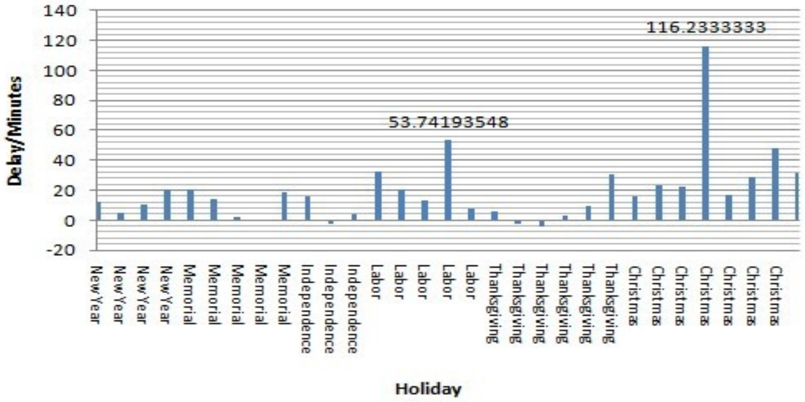


FIGURE 4.1B: HOLIDAY EFFECT OF 2012

So combing this character with the usage of weights of knots in smoothing splines can make a better fit of the data and evaluate the data more precisely.

The functions can be expressed as the following formulas:

$$S(d) = \sum_{j=0}^{m} \beta_j D^j + \sum_{j=1}^{k} \lambda_j (d- \xi_k)^m{}_+$$

where the notation

$$(d-\xi_k)_+ = \begin{cases} d - \xi_k, d > \xi_k \\ 0, otherwise \end{cases}$$

And the penalty function is known as the following formula:

$$PSS=\sum_{d=1}^{n}(y_d - S(d))^2 + \lambda \int (S''(d))^2 dx$$

where λ is a tuning parameter. y_d denotes the average day delay and can be calculated by

$$y_d = \frac{\sum_t \sum_i y_{dti}}{\sum_t n_{dt}} \qquad d=1, 2, 3, \ldots, n$$

where y_{dti} denotes the observed delay of flight i, at time t and on day d; n_{dt} denotes number of flights at time t on day d. For PSS function, if λ → 0, it means there is no penalty and the spline can have a very close fit, but the result curve could be very noisy as it follows every detail in the data. If λ →∞, the effect of the spline is the opposite. The spline can be very smoothly, but may be a bad fit.

Instead of normal smoothing function usually with same weight at each knot, here we define the Weight Index using the following formula,

$$\omega_d = \begin{cases} \frac{y_{dti}}{|yd|} * 100 & , y_{dti} > 0 \\ 1 & , y_{dti} \leq 0 \end{cases}$$

4.1.2 Time-Schedule Function

Time-Schedule effect reflects the pattern of different hour and even different minute. The pattern can be caused by airport condition whether it is large enough to handle a burst of flights departure at similar time.

For time-schedule effect, we still employ weighted smoothing splines to fit the data. The only thing need to pay attention is all delays caused by different factors will be added together to get the final delay. Hence when we use data for time-schedule effect should remove the delay caused by weather and holiday effects. The formulas are as following:

$$y'_{dti} = y_{dti} - S(d) \qquad \forall\, d, t, i$$

$$PSS = \sum_{t=00:00}^{24:00} (y_t - S(t))^2 + \lambda \int (S''(t))^2 dt$$

$$y_t = \frac{\sum_{d=1}^{365} \sum_i y'_{dti}}{\sum_{d=1}^{365} n_{dt}}$$

y_t denotes the average delay of all flights departure at time t.

4.1.3 Random Factor Function

The random factors are those don't belong to the weather&holiday or the time-schedule effect. Those factors stem from random situation like luggage problem, terrorist attack, mechanic problem or other emergency situations. In order to model the random factors' delay, we utilize an ARIMA model to capture the trend.

It follows the formula below:

$$\begin{cases} \Phi(L)\Delta^d y_t = c + \Theta(L)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2, E(\varepsilon_t, \varepsilon_s) = 0, s \neq t(1) \\ Ey_s\varepsilon_t = 0, \forall\, s < t \end{cases}$$

where $\Delta^d = (1 - L)^d$,

$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^P$, which is the auto regression (AR) coefficient polynomials;

$\Theta(\mathrm{L}) = 1 + \theta_1 \mathrm{L} + \theta_2 \mathrm{L}^2 + \cdots + \theta_q \mathrm{L}^q$, which is the moving average (MA) coefficients polynomials;

$\{\varepsilon_t\}$ is zero mean white noise sequence

ARIMA modeling steps has been introduced in Chapter 3.

This basic model is comprehensive for long term general prediction of all airports and airlines. Since it only using historical delay data to make a general prediction, **which means two different flights depart at similar time will have similar delay predictions**. The reason why this happened is $\omega(dh)\ and\ \delta(t)$ on the same day will be the same. The only difference is the random factor $\in_n$ will be different for each flight, but not big difference. The performances of this model will be shown in Data Testing &Model Evaluation section. But nowadays what customers need is a real time prediction of a specific flight which concerns up to data conditions including (current weather, aviation system of airports, last late aircraft and etc.). Besides, customers focus on arrival delays more than departure delays actually. Hence, as to the two limits of the basic model, now we will introduce our final real time model based on modification of the basic model.

4.2 Improved Real Time Arrival Prediction Model

To overcome the defects of the basic long term prediction model, here we focus on building a model which can combine real time data to give customers the latest arrival status information about their flights. The difference between the new model and the basic one is we build models for each flight at a specific airport instead of considering all airlines have similar delay model. The reason why we use every flight historical data to train each model is that we found out each flight/aircraft has its own delay pattern which is different from the other flights. This is due to each flight/airline has different time schedules, flight crews, airport conditions, weather influence and this characteristic has been shown in Data Exploration Chapter.

Hence in order to obtain accurate models, we have to train each model with respective real time data to get model parameters. Now we will introduce this model with more details.

In this real time model, we group all factors into two main categories by using all real time data. The model structure is shown in Figure 4.2A Improved Real Time Prediction Model Structure.

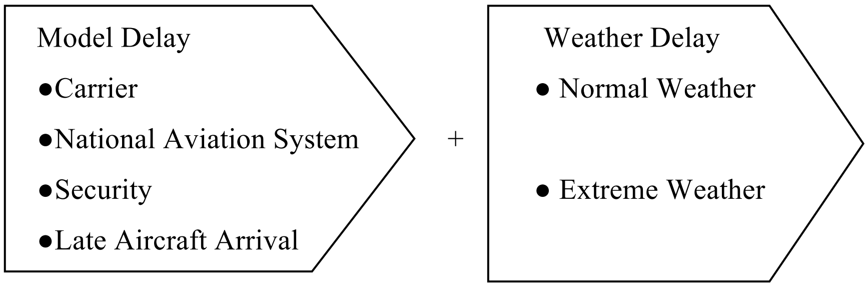


Figure 4.2A: Improved Real Time Prediction Model Structure

4.2.1Model Delay Function

The first category is Model Delay which considers carrier, national aviation system, security and late aircraft arrival as factors will contribute to delays. As we stated in previous chapter, carrier factor means delay due to airline's control such as crew problems, maintenance, cleaning, baggage loading or fueling, etc. National Aviation System factor shows delay because of airport operations, heavy traffic volume, air traffic control and etc. Security factor denotes delays or cancellations caused by evacuation of a terminal or re-boarding problems, screening equipment problems and etc. Late Aircraft Arrival factor illustrates delay caused by a previous flight with same aircraft arrived late. All historical data for each factor can be obtained from The Bureau of Transportation Statistics (BTS), which will be shown in Chapter 5 Data Application & Testing.

I. Mathematic Function Selection

Among numerous fitting functions we have studied during this project, here we still choose smoothing spline function as Model Delay Function.

For each flight, we use the following formulas:

$$S(f) = \sum_{j=0}^{m} \beta_j D^j + \sum_{j=1}^{k} \lambda_j (f - \xi_k)^m{}_+$$

where f means flight arrival delay in minutes due to Model Delay (Weather Delay excluded) sequenced by time of each flight and the notation

$$(f - \xi_k)_+ = \begin{cases} f - \xi_k, f > \xi_k \\ 0, otherwise \end{cases}$$

And the penalty function is known as the following formula:

$$PSS = \sum_{f=1}^{n} (y_f - S(f))^2 + \lambda \int (S''(f))^2 dx$$

where λ is a tuning parameter. Y_f denotes the average delay of a flight and can be calculated by

$$y_f = \frac{\sum_t y_{ti}}{n_{ti}}$$

where y_{ti} denotes the observed delay of flight i, at time t; n_{ti} denotes total number of flights of flight i until time t .

4.2.2 Weather Delay Function

The second part is Weather Delay. In this part, we consider normal weather and extreme weather influence. As proved in Data Exploration Chapter, weather factor actually contributes almost 30% to 50% delays among all kinds of factors. Therefore, the accuracy of a weather delay prediction model cannot be more important.

I. Weather Indicators Selection

To describe weather in a location, usually we need several indicators to illustrate the situation. These indicators include temperature, wind, visibility, precipitation and sky conditions (cloud thickness, snow, rain, fog, etc.).

Among these indicators, we need to figure out main factors which have an influence on flight delays. Here we use principal component analysis and factor analysis (PCFA) method [22] to get main factors which contribute to flight delays. By applying weather data to delay data due to weather part, we get three main factors in our model, which are **wind, visibility and sky conditions**.

● Wind Indicator

Wind elements include average daily speed, current wind speed, wind direction vector, gust speed, fastest 5-second wind speed, fastest 2-minute wind speed and etc. Among so many elements, we found out average daily speed doesn't have an obvious effect on flight arrival delays. Instead, current wind speed shows a strong correlation on flight arrival delays. Also, if the gust speed is much higher when the flight is going to land, it will prevent a timely and safely landing of the flight. Beside the wind speed influence, wind direction can also shows impact on flight arrival delays since a varying wind direction will put a flight at a risk. In conclusion, we utilize elements like current wind speed, gust speed and wind direction as our wind indicator.

● Visibility

Visibility is the ability to see an object in the atmosphere. In terms of the weather, visibility is the greatest horizontal distance, at which selected objects can be seen, identified, and/or measured with instrumentation.[23]

For a safe flying, the pilot needs a minimum amount of visibility for landing at the airport. Clear clean air has a better visibility than air polluted with dust or other particles. This depends on a number of factors which are all weather related. Study shows there is no difference of visibility or transparency of air between day and night. Hence, sun or moonlight does not alter the transparency of the air.[24]

● Sky Conditions

Except the first two indicators wind and visibility, we also find out there is a correlation between weather delay and sky conditions, which means different sky conditions will have a different influence on flying.

There are a lot of sky conditions can influence flying:

◆ Cloud Thickness: There are several standards to summarize cloud thickness. In this thesis, we use METAR, which is a format for reporting weather. Raw METAR is the most common format in the world for the transmission of observational weather data. It is highly standardized through the International Civil Aviation Organization (ICAO), which allows it to be understood throughout most of the world.[25]

In METAR, we category cloud thickness into 8 kinds, which are show in Table 4.2A METAR Cloud Thickness Categories [26]:

Abbreviation	**Meaning**
SKC	"No cloud/Sky clear" used worldwide but in North America is used to indicate a human generated report[12][13]
CLR	"No clouds below 12,000 ft (3,700 m) (U.S.) or 10,000 ft (3,000 m) (Canada)", used mainly within North America and indicates a station that is at least partly automated[12][13]
NSC	"No (nil) significant cloud", *i.e.*, none below 5,000 ft (1,500 m) and no TCU or CB. Not used in North America.
FEW	"Few" = 1–2 oktas
SCT	"Scattered" = 3–4 oktas
BKN	"Broken" = 5–7 oktas
OVC	"Overcast" = 8 oktas, *i.e.*, full cloud coverage
VV	Clouds cannot be seen because of fog or heavy precipitation, so vertical visibility is given instead.

TABLE 4.2A: METAR CLOUD THICKNESS CATEGORIES

- Precipitation: Rain or snow will reduce visibility. Of course it depends a bit on how heavy the precipitation, drop or snow flake size and the intensity are. A light drizzle will not hinder VFR operations (although commercial operations usually will have higher limits, see part 91 vs 125/135) but heavy precipitation in Cb or TCu can reduce visibility to 100 meters or even less accompanied with effects like wind shear and turbulence.[27]

- Fog/Mist: People often get confused between fog and mist. Fog means visibility is less than 1000 meters and mist is visibility between 1000 and 5000 meters. But both fog and mist have their origins in light suspended cloud droplets with almost 100% relative humidity and an abundance of condensation nuclei for the condensation process to start.[28] Hence, both fog and mist will have influences on weather delays.

- Haze: It is traditionally an atmospheric phenomenon where dust, smoke and other dry particles obscure the clarity of the sky [29].When visibility is reduced to 5000 meters or less *by the presence of dust particles* it is called haze. When there is a serious haze, it will influence a flight's landing.

- Sand Storm: When dust or sand particles are blown off and visibility reduces to less than 1000 meters it is referred to as a dust or sand storm, with altitudes usually not higher than around 150 - 200 ft.

- Other Extreme Weather Phenomenon: Tornado, Hurricane, Thunderstorm, Volcanic Ash and etc. All these extreme weather phenomena will have a significant impact or even prevention on flight landings.

II. Mathematic Function Selection

We use Multiple Linear Regression for weather delay analysis in this thesis. There are three predictors in our model which are wind speed, visibility and sky conditions. And we use delay time due to weather factor (in minutes) as the dependant variable. So our multiple linear regression function has the following form:

$$y_i=\beta_1 w_i + \beta_2 v_i + \beta_3 s_i + \varepsilon_i \quad i = 1, \ldots, n,$$

Where w_i refers to wind speed, v_i represents visibility and s_i stands for sky conditions. And i is on behalf of a flight number, y_i means weather delay of flight i. By applying all historical weather delay data sets (obtained from BTS and weather history data) to this linear regression and minimize the likelihood value, we can get reasonable coefficients which are β_1, β_2, β_3 in our model.

CHAPTER 5 DATA APPLICATION & TESTING

Up to now, we have described two models with different methods, one for long term general prediction and one for up to date real time prediction of a specific flight. Here in this chapter, we will show how we apply data sets to those two models and produce model parameters. Besides, we will also give results evaluation of each function we have obtained. The general evaluation of both models will be given in Chapter 6.

5.1 Long Term General Departure Prediction Model

In first model, we select San Francisco International Airport and use 3 years data (2010~2012) collected from The Bureau of Transportation Statistics (BTS). There are more than 10 carriers and we choose one of the biggest, American Airline to specifically obtain an accurate and high qualified model. Hence the other airports and airlines can also apply into our model according to respective data.

1) Data Process

There are more than 30,000 records of data within 3 years for American Airline. All records are stored chronologically in database and in order to study and compare every year's pattern, we delete February 29, 2012 since there are 366 days in 2012 and only 365 days in both 2010 and 2011. Every record consists of four components in our database: Data, Flight_No, Scheduled_time and Delay. 33348 American Airline flights departed from San Francisco International Airport in 3 years in total, which are around 30 flights each day. And we will use data in 2013 to test our model, but also will delete specific days with emergency incidents happened such as the crash happened in July 13, 2013 which has caused several fatalities and influenced the whole airport normal function. Some news can be found on that day [30]:

A Boeing 777 airliner, operated by Asiana Airlines, crashed on landing at San Francisco International Airport on Saturday. The San Francisco International Airport (SFO) was closed as of Jul 06 at 01:10 PM PDT. The date/time when the airport is expected *to reopen is not known."*

2) ***Weather&Holiday Effect Modeling***

a) Weather Effect

After gathering all records from 2010 to 2012, we plot the average delay trend in R using total delay everyday divided by departure numbers of flights (Fig. 5.1A). The x-axis gives the day number, which means we calculate different days using an increasing sequence from day 1 (01.01.2010) to day 1095 (31.12.2012). The y-axis denotes the average delay in minutes.

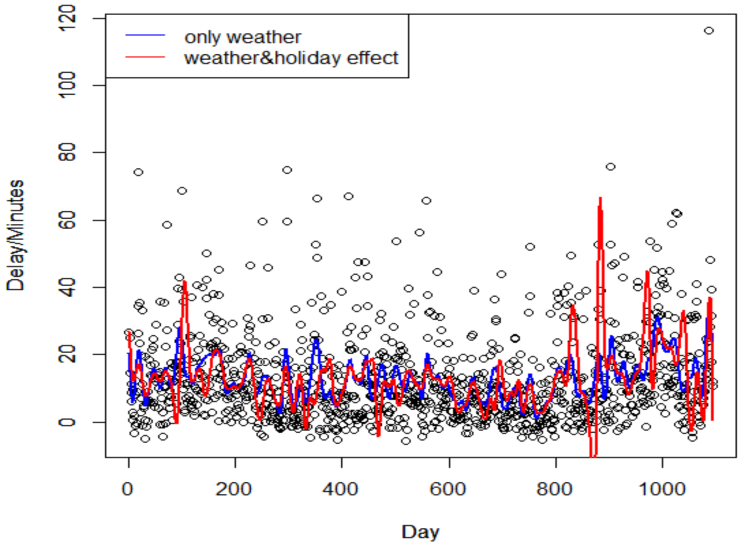


FIGURE 5.1A: DAILY AVERAGE DAY TREND

After plotting the basic data in R, we use method smooth.spline in package stats to fit the data and get a general trend, which is represented as the blue line in Figure 5.1A with smoothing parameter spar = 0.24. A spar parameter is choose from a λ by minimizing PSS=$\sum_{d=1}^{n}(y_d - S(d))^2 + \lambda \int (S''(d))^2 dx$, where λ = r * 256^(3*spar – 1)[31]. The most suitable spar should prevent data over fit and at the same time yield the minimum deviation.

b) Holiday Effect

Here we probably found improper in this trend, since in method smooth.spline, it sets the weights of all knots as 1 which means every point has the same importance. But the real data shows obviously some days have much larger delay than the other days. For example, December 23, 2012, the average delay is 116 minutes which is much higher than day average delay-12.8 minutes in 3 years. Figure 5.1B Holiday Effect depicts the main holiday trend of 2012, where we can find the average delay during holiday is much higher than normal days. In order to capture this character and also devoid the inaccuracy of method 1, we apply different priorities on holiday data by setting high values in weight parameter in smooth.spline method. The weight value of each data point is calculated by the formula we defined before in section II. The red line in Fig. 5.1A shows the new trend of our data which shows a better fit on days with holiday effect. Here we choose spar=0.14 instead of 0.24.

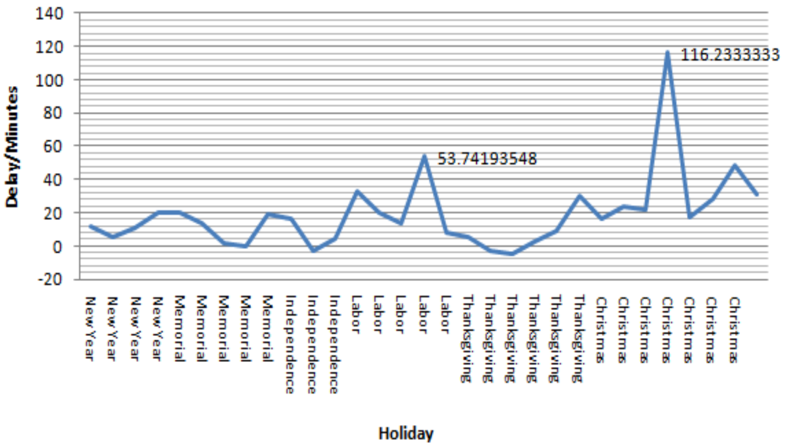


FIGURE 5.1B: HOLIDAY EFFECT

c) Forecast with Weather&Holiday Trend

In order to test the reliability of the predictions of weather &holiday effect, we generate the function from the training set which is 80% of data in a day in 3 years, and we plot the 20% corresponding data left which is called holdout data on the same graph. In Fig. 5.1C, the x axis means the corresponding day in 3 years and the y axis denotes the average day delay calculated by data in holdout set. In order to see clearly, we only plot part of the days nearly from day 800 to day 1100. We can see the spline function generated by our model effectively captured useful information in the training data and fit the holdout data perfectly.

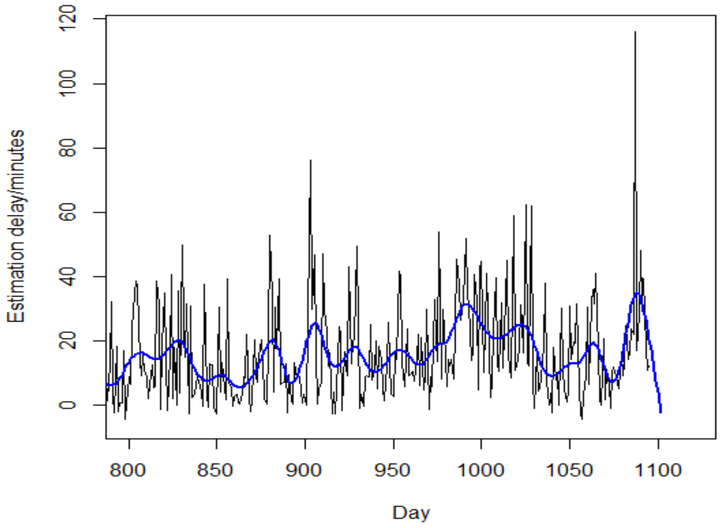


FIGURE 5.1C: ESTIMATION OF HOLDOUT SET

3) Time-Schedule Effect Modeling

In order to model time-schedule effect, we need to remove the first factor--- weather & holiday effect. Fig. 5.1D depicts the data distribution after removing first factor effect. The x axis represents day and the y axis means day average delay after subtracting delay caused by first factor. The current day average delay is fluctuating above or below 0 which also reflects a good fit of first factor.

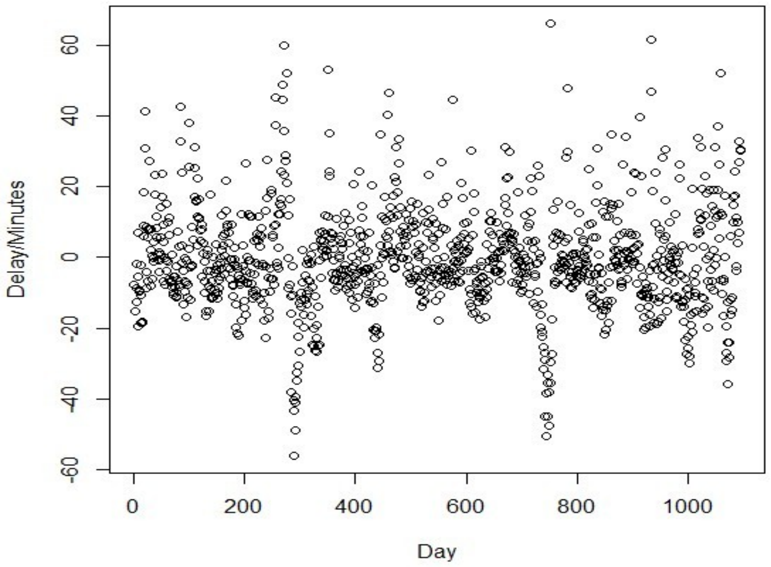


FIGURE 5.1D: DATA DISTRIBUTION AFTER REMOVING FIRST FACTOR

After achieving data without first factor, we rearrange the data according to time-schedule sequence, which means data records in three years are sequenced from 00:00 to 23:59. Now we can get the average delay for each time-schedule data point by formula (9) described in section II. Then we apply a similar smoothing spline method to estimate time-schedule model. Fig. 5.1E presents a fitted smoothing spline with spar = 0.59. The x axis is the scheduled departure time in minutes calculated from 00:00 to 23:59, and the y axis is the average delay in minutes. Please note that between 1:00 to 5:59 and between 21:00 to 21:59, there is no flight was scheduled to depart in those three years.

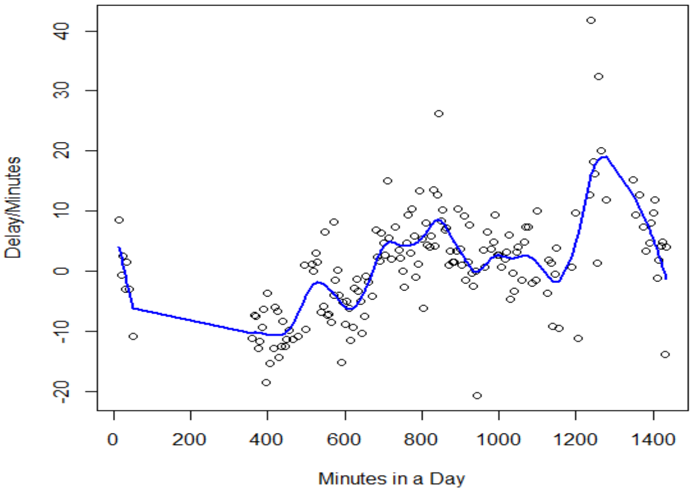


FIGURE 5.1E: TIME-SCHEDULE SPLINE

We can tell from the trend that the average delay gradually increase as time goes by in day time and evening but decrease when late in night. The fluctuation and the increasing tread sufficiently demonstrate the significance of time-schedule effect.

4) Random Factors Effect

a) Difference Operation

After removing the first two factors, we can plot the data distribution by Fig. 5.1F. The x axis stands for each flight ranged by its corresponding time schedule in three years. And the y axis means flight delay for each flight in minutes after removing the first two factors. This means the original delay of each flight needs to minus delay caused by first factor (weather & holiday effect) firstly and minus second factor (time-schedule effect) secondly.

From the figure, we can see a large percent of data are around 0 minutes, only few data points are around or above 1000 minutes. These points actually can be regarded as outliers caused by specific incidents like severe weather condition, terrorist threatens or other uncommon reasons. But here we will not delete them, they will still contribute some effect on our factors especially when we calculate holiday effect.

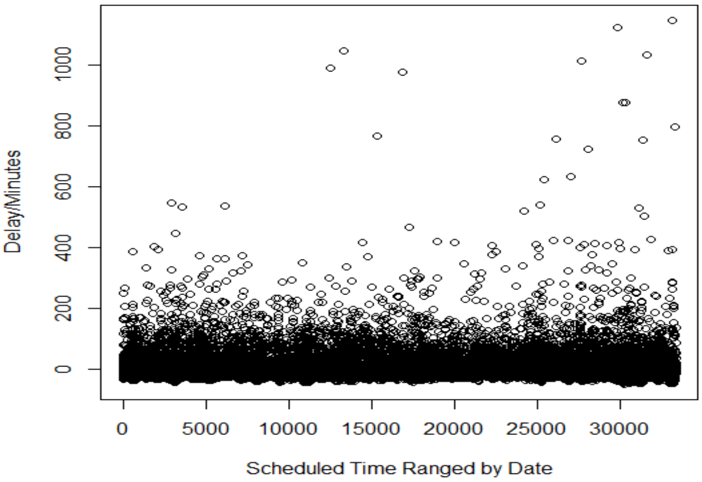


FIGURE 5.1F: DATA DISTRIBUTION AFTER FIRST TWO FACTORS

By drawing the ACF in top graph in Fig.5.1G, we can see that it does not have tail off characteristic before difference which means it is not a stationary time series. Hence, we need to figure out how many difference times we need to use to make it stationary. After the first time of difference operation, we get the ACF and PACF pattern in the middle and bottom of Figure 3.4 b, which shows a tail off/cut off character in ACF after Lag =2 and PACF is negative.

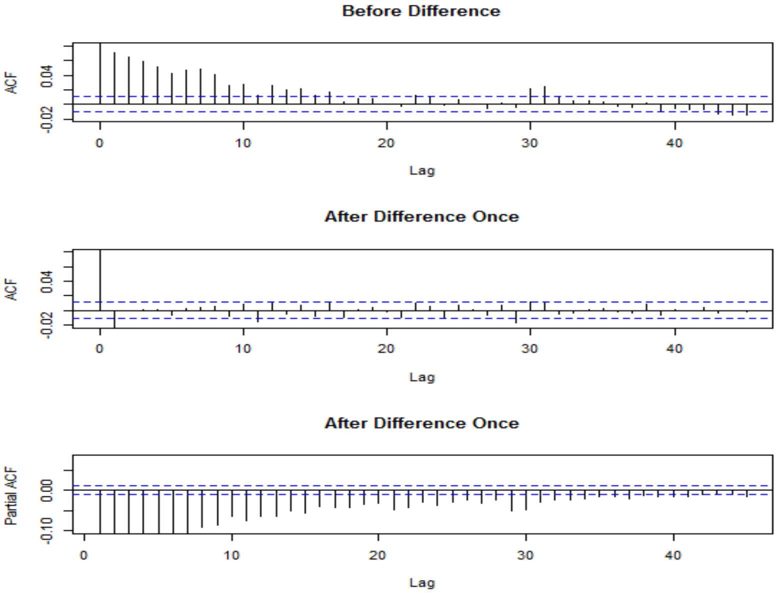


Figure 5.1G: ACF&PACF Before and After Difference

b) Stationary Test

By using Augmented Dickey-Fuller Test (Table 5.1A), we can know the time series is stationary:

Table 5.1A: Augmented Dickey-Fuller Test

```
data:    random factors after 1st difference
Dickey-Fuller = -53.0158, Lag order = 32, p-value = 0.01
alternative hypothesis: stationary
```

c) ARIMA Model Parameters

After testing different combinations of p,q in ARIMA(p,1,q), we found when p=2 and q=2 yields to the finest ARIMA model which has the least AIC value compare to other combinations. The detail of the coefficients is as following (Table 5.1B):

Table 5.1B: ARIMA Parameters

```
arima(x = y3, order = c(2, 1, 2))
Coefficients:
            ar1       ar2        ma1       ma2
         0.6953    0.0236    -1.6597    0.6598
s.e.     0.0158    0.0057     0.0149    0.0149
sigma^2 estimated as 1852:    log likelihood = -172776.5,
                                    aic = 345563
```

From the table the two coefficients of AR model is 0.6953 and 0.0236 with standard error 0.0158 and 0.0057 respectively. And the two coefficients of MA model is -1.6597 and 0.6598 with same stand error 0.0149. The formula yielded is as following:

$Y_t= 0.6953y_{t-1} + 0.0236y_{t-2} -1.6597e_{t-1} + 0.6598e_{t-2} + e_t$

d) ARIMA Model Validation

Here we use Box-Ljung test to see whether the residuals of ARIMA model is random or not. The p-value in Table 5.1C shows the model is valid.

Table 5.1C: Box-Ljung Test

```
        Box-Ljung test

data:   y3.fit$resid
X-squared = 141.7454, df = 6, p-value < 2.2e-16
```

e) Forecast with ARIMA

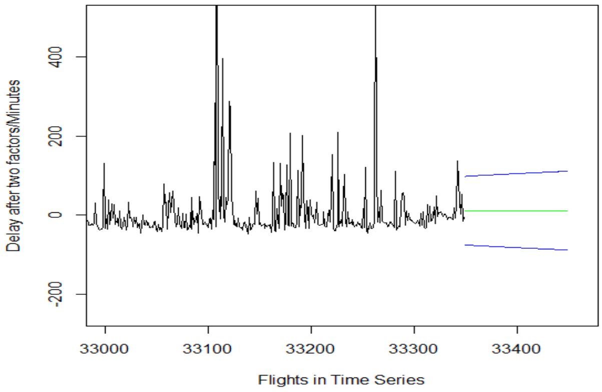


FIGURE 5.1H: FORECAST RANDOM FACTOR EFFECT

In Fig.5.1H, the x axis stands for each flight ranged by its corresponding time schedule in three years. The y axis represents each flight delay after removing first two factors. The black data points shows the residuals left after subtracting weather & holiday effect and time-schedule effect which also stands for random factor values. The green line is generated to predict random factor values in the following 100 days. The two blue lines shows the range of our prediction which is calculated by predictions ±2* standard error of predictions. A mean value of predictions can be inferred from the plot. Please note the green line is not totally horizontal. Every data point on the green line varies a little bit from each other.

5.2 Improved Real Time Arrival Prediction Model

In this model, we still select American Airline at San Francisco International Airport. Since our goal is to build a real time prediction models for each flight (here flight means flight with same flight number), for Model Delay Function, we utilize each flight's latest three months historical arrival model delay data(does not include weather delay) from The Bureau of Transportation Statistics (BTS) to produce the model. At the same time, for Weather Delay Function, we apply 2010's historical weather data to flight arrival weather delay data in 2010 at San Francisco International Airport to obtain coefficients in our multiple regression model. The model building methods can be used in any other airlines and airports.

1) Data Process

a) Model Delay Function

Concerned about there are 32 American Airline flights at San Francisco International Airport at present (may change in the future), we fetched arrival delay data for each flight from December 1, 2013 to February 28, 2014. Those 32 flight numbers are shown in the following Table 5.2A: Flight Number in American Airline at SFO Airport. In total, there are around 3000 records for these flights. By subtracting arrival weather delay part (in minutes) from total arrival delay part (in minutes), we obtained model delay part (in minutes). Hence, in our database, each record consists of Date, Flight_No and Arrival_Model_Delay.

Flight No.	
59	69
85	108
122	167
177	179
193	197
203	219
275	323
329	399
1105	1399
1427	1521
1524	1536
1585	1658
1677	2245
2356	2411
2455	2456
2457	2465

Table 5.2A: Flight Number in American Airline at SFO Airport

- Fitting Method of Model Delay

After storing data in our database, now we apply smoothing spline function in R, the main codes are shown in Table 5.2B: Model_Delay_Fitting_Method

```
 #Read in arrival_model_delay data of a specific flight
data<-read.csv('Arrival_Model_Delay_of_FlightNo.csv')
names(data)<-c('ModelDelay')
y<-data$ ModelDelay
n<- 1:length(y)

#Plot out arrival_model_delay of the specific flight
plot(y~n,xlab="Day",ylab="Model Delay/Minutes")

#Applying smoothing spline function to the data
sp<-smooth.spline(n,y)

#Show fitting parameters
sp

#Draw the fitting function on the same graph
lines(sp,col=2,lwd=2)

#Delay prediction for the next arrival
x_left<-length(y)
x_right<-length(y)+1
x_prediction_interval<-c(x_left, x_right)
prediction<-predict(sp, x_prediction_interva)

#Get prediction value
prediction$y

# Write prediction values to a csv file
write.table(prediction$y, file = "prediction_of_FlightNo.csv",
sep = ",", col.names = NA,qmethod = "double")
```

TABLE 5.2B: MODEL_DELAY_FITTING_METHOD

- Case Study

In this section, we take three months data of Flight No.59 of American Airline at San Francisco International Airport as an example to illustrate how Model Delay Function will be generated. After importing model delay data of Flight No.59 into R, we carry out plotting function to plot all model delay values from December 1, 2013 to February 28, 2014. The graph is presented by Figure 5.2A: Model Delay Scatter Diagram of Flight No.59

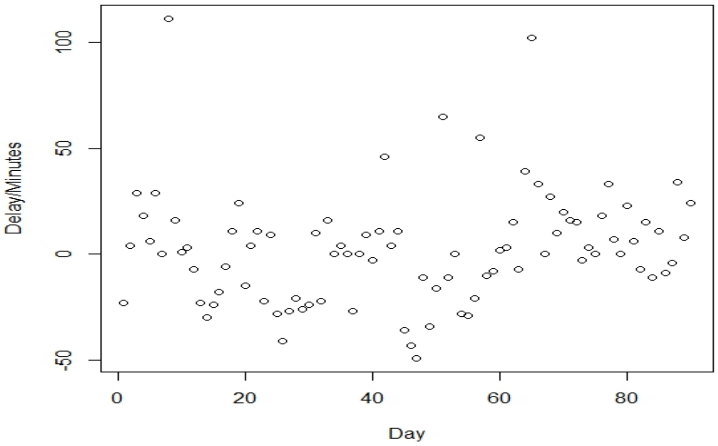


FIGURE 5.2A: MODEL DELAY SCATTER DIAGRAM OF FLIGHT NO.59

After plotting the data, we apply smooth.spline function to conduct the fitting process by using 90% of the whole data which are from December 1, 2013 to February 18 and we will using the 10% holdout set for the prediction. The fitting function is shown in Figure 5.2B: Smoothing Spline Fitting Function of Flight No.59 and the fitting parameters we obtained are shown in Table 5.2C: Smoothing Spline Fitting Parameters of Flight No.59.

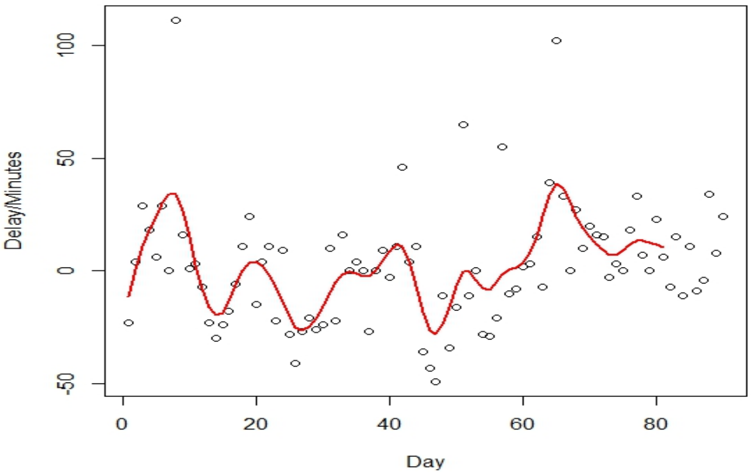


FIGURE 5.2 B: SMOOTHING SPLINE FITTING FUNCTION OF FLIGHT NO.59

```
Call:
smooth.spline(x = n1, y = y1)
Smoothing Parameter    spar= 0.4659227
lambda= 1.195225e-05 (12 iterations)
Equivalent Degrees of Freedom (Df): 18.97435
Penalized Criterion: 31602.6
GCV: 665.373
```

TABLE 5.2C: SMOOTHING SPLINE FITTING PARAMETERS OF FLIGHT NO.59.

After getting the fitting function by using smooth.spline in R, we can easily do prediction for the left 10% holdout set which are from February 19 to February 28.

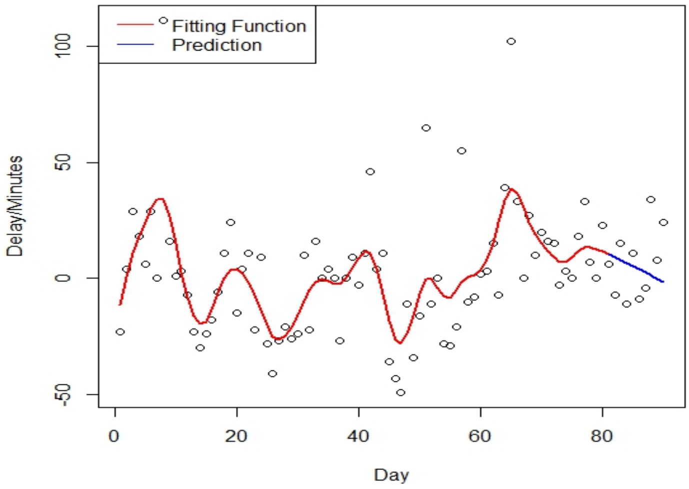


FIGURE 5.2C: PREDICTION OF MODEL DELAY

Figure 5.2C: Prediction of Model Delay has shown our predictions for the following 10 days. The blue prediction curve has apparently shown a good prediction for Model Delay already for Flight No.59. The following Table 5.2D: Real Model Delay Data vs. Predictions of Flight No.59 gives us a more accurate and straight way to evaluate our fitting function.

Date	Real Data/Min	Prediction/Min	Difference/Min
19/02/2014	6	10.4981002	-4.498100
20/02/2014	-7	9.1702696	-16.170270
21/02/2014	15	7.8424391	7.157561
22/02/2014	-11	6.5146085	-17.514608
23/02/2014	11	5.1867779	5.813222
24/02/2014	-9	3.8589473	-12.858947
25/02/2014	-4	2.5311168	-6.531117
26/02/2014	34	1.2032862	32.796714
27/02/2014	8	-0.1245444	8.124544
28/02/2014	24	-1.4523750	25.452375
Mean	6.7	4.522863	2.177137

TABLE 5.2D: REAL MODEL DELAY DATA VS. PREDICTIONS OF FLIGHT NO.59

- Result Analysis

From results in Table 5.2D, the absolute differences between real data and predictions are ranging from 4.5 minutes to 32.7 minutes. If we assume the error range is 30 minutes, then the prediction accurate rate is 90%. At the same time the mean value of differences is only around 2.2 minutes which also reflects the stability and tolerance of our prediction. **Besides**, here we do predictions for the next ten days, which is a long prediction range in flight delay prediction area. In real system, we will use **the latest real time real data to produce our fitting function** due to the reason that fitting curve is changing all the time when new data added. **Meanwhile** we only do predictions for the next **two days**, in which we can guarantee even higher prediction accuracy.

b) Weather Delay Function

In order to obtain the relationship between weather effect and flight delay, in this section, we apply historical weather data in 2010 of San Francisco International Airport to the corresponding arrival weather delay data in 2012.

Here, as regards to historical weather data, we fetched from the website *www.wunderground.com* which provides current and historical weather data inquiry. As we analyzed in previous chapter, we downloaded weather data consists of wind speed, visibility and sky condition. And we converted visibility and sky conditions into mathematic numbers according to our defined methods. The conversion methods are presented in the following Table 5.2E: Visibility Conversion Method and Table 5.2F: Sky Conditions Conversion Method. Since there are around 12,000 flights of American Airline arrived at San Francisco International Airport in 2010, there are around corresponding weather records in the database.

Visibility	Converted Points
(0,5]	10
(5,7]	8
(7,8]	6
(8,9]	4
(9,10)	2
10	0

TABLE 5.2E: VISIBILITY CONVERSION METHOD

Sky Conditions	Converted Points
Hurricane/Tornado	>=50
Heavy Rain/Heavy Snow/Heavy Fog/Heavy Drizzle/Heavy Mist/Heavy Haze/Heavy Hail/Thunderstorms/Overcast	10
Moderate Rain/Moderate Snow/Moderate Fog/Moderate Drizzle/Moderate Mist/Moderate Hail/Broken Clouds	8
Light Rain/Light Snow/Light Fog/Light Drizzle/Light Mist/ Light Hail/Scattered Clouds	4
Else	0

TABLE 5.2F: SKY CONDITIONS CONVERSION METHOD

With respect to arrival weather delay data in 2012, we can acquire them by subtracting model delay minutes from total delay minutes. After the above steps, part of the records in our database are shown in the following Table 5.2G: Multiple Regression Data of Weather Delay

TABLE 5.2G: MULTIPLE REGRESSION DATA OF WEATHER DELAY

Weather Delay/Mins	Wind Speed/Knots	Visibility Points	Sky Condtion Points
226	27.21376	8	10
219	23.32608	8	10
197	19.4384	8	10
191	30.90706	6	8
188	19.4384	8	8
159	21.38224	10	10
149	23.32608	4	8
145	23.32608	4	8
142	21.38224	4	4
119	19.4384	4	4
116	15.55072	8	10
114	23.90923	2	8
114	23.90923	2	8
112	19.4384	2	8
107	17.49456	2	8
104	16.91141	2	8
86	11.66304	2	10
85	12.05181	10	10

- Fitting Method of Weather Delay

After the preparation of all related analysis data in Weather Delay Function, we apply Multiple Linear Regression to our model. The coefficients of the regression will show the relationships between Wind Speed, Visibility, Sky Conditions and Weather Delay.

- Case Study

Due to the reason that all original data from the BTS are reported and analyzed by airlines' manual records, there are some inevitable errors exist in those records, which means all weather parameters we found for that flight should not lead to such a large weather delay minutes. Hence, before we apply the regression function, we need to examine and verify some unreasonable weather data. By filter all data records for Weather Delay Function, we found the following unreasonable ones in Table 5.2H: Unreasonable Weather Delay Records

:

TABLE 5.2H: UNREASONABLE WEATHER DELAY RECORDS

Date	Flight No.	Delay/Mins	DelayWeather/Mins	Sky Condition	WindSpeed/Knots	Visibility(km)
04/07/2010	1309	213	213	Partly cloudy	14.96760257	16
03/10/2010	1575	22	8	Clear	17.10583151	16
12/12/2010	451	27	8	Clear	14.96760257	16
06/27/2010	451	37	1	Clear	13.99568033	16
07/06/2010	2281	104	1	Clear	13.41252698	16

From Table 5.2H we can find out that on July 4, 2010, Flight 1309 delayed 213 minutes in total, which is all due to Delay Weather factor. Theoretically, there must be a serious weather condition around the arrival time of Flight 1309. However, after fetching the weather parameters during that period, we found the wind speed is low with stable wind direction, the visibility is totally in good condition and there is no rain, no snow and no other extreme weather conditions. The sky condition is partly cloudy which has not been proved will lead to such a long arrival delay situation. Hence, we consider this record needs to be deleted when training our model by using these data.

The following 4 records which have same sky condition and visibility, the only difference is wind speed. Flight 1575 has much higher wind speed than the other three flights. It should have higher Delay Weather value than Flight No.451 on December 12, 2010. And as the last two records, the weather conditions lead to only 1 minute delay which is too little time to be accurate for fitting functions. In order to have a better model function, we abandon those small Delay Weather values.

After filtering all records we fetched in 2010, now we can apply multiple linear regression method in Excel or R to acquire model parameters. The following Table 5.2I Multiple Linear

Regression Results of Weather Delay Function shows all parameters and results.

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.949519053
R Square	0.901586432
Adjusted R Square	0.901559606
Standard Error	6.655528791
Observations	11010

ANOVA

	df	SS	MS	F	Significance F
Regression	3	4466291.142	1488764	33609.39092	0
Residual	11006	487522.4748	44.29606		
Total	11009	4953813.616			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
ntercept	-0.39290192	0.066517714	-5.90673	3.59297E-09	-0.523288582	-0.262515265	-0.523288582	-0.262515265
WindSpeed	3.319389823	0.024575518	135.069	0	3.271217396	3.36756225	3.271217396	3.36756225
Visibility	8.308365516	0.078569962	105.7448	0	8.154354288	8.462376744	8.154354288	8.462376744
SkyConditions	0.764536517	0.050899816	15.02042	1.70568E-50	0.664763742	0.864309292	0.664763742	0.864309292

TABLE 5.2I: MULTIPLE LINEAR REGRESSION RESULTS OF WEATHER DELAY FUNCTION

- Result Analysis

Firstly, from the results, we can see coefficients of Wind Speed, Visibility and Sky Conditions are 3.319389823, 8.308365516 and 0.764536517 respectively. Intercept of our equation is -0.39290192. Hence, our multiple linear regression function can be expressed as following formula:

$$y_i=3.319389823 * W_i + 8.308365516 * V_i + 0.764536517 * S_i - 0.39290192 \quad i = 1, \ldots, n$$

Secondly, we can analyze the goodness of this function[32].

◆ Overall Regression's Accuracy

R Square – As we know, this is the most significant data of the results. R Square indicates how well the regression line approximates the real data. This number tells you how much of the output variable's variance is explained by the input variables' variance. Ideally, we would like to see this at least 0.6 (60%) or 0.7 (70%). In our model, **R Square is 0.901586432**, which means 90% variance of weather delay minutes can be explained by Wind Speed, Visibility and Sky Conditions.

Adjusted R Square – This is used most often when we illustrate the accuracy of the regression function. Adjusted R Square is more conservative than R Square because it is always less than R Square. Beside, when new input variables are added to the Regression analysis, only when the new input variable makes the regression function more accurate, Adjusted R Square will increase. If new input variable didn't make the regression function more accurate, Adjusted R Square will not increase. In contrast, R Square always goes up when a new variable is added; no matter the new input variable improves the Regression equation's accuracy.

◆ *Whether Results Are By Chance Or Not*

Significance of F – This evaluates the probability that the Regression results could have been produced by chance. A small Significance of F proves the validity of the Regression results. Given an example, if Significance of F = 0.04, this means there is only a 4% chance that the Regression result is just a chance occurrence. Hence, the function we obtained is apparently not by chance since Significance of F almost equals 0.

◆ *Individual Regression Coefficient Accuracy*

P-value – The P-Values of each parameter indicates the likelihood that they are real results and have not occurred by chance. The smaller the P-Value is, the larger the likelihood that the coefficient or Y-Intercept is valid. As an example, a P-Value of 0.02 for a regression coefficient indicates that there is only a 2% chance that the result occurred only by chance. Again in our results, the P-Values of four variables are all less than 0.01 which indicate these individual regression coefficients are accurate and do not occur as an occurrence.

CHAPTER 6 MODEL EVALUATION

Even though we have applied datasets for both prediction models and evaluated functions of each factor in Chapter 5, now we will give a more comprehensive evaluation for whole models.

6.1 Evaluation of General Long Term Departure Prediction Model

In this section, we mainly focus on forecast the probability that the delay of a flight is shorter than fixed minutes. We are still training our dataset using 80% data and predict using the left 20% holdout set from 2010 and 2012. Besides, we set the confidence interval (CI) as 80%.

TABLE 6.1A: GENERAL LONG TERM DEPARTURE PREDICTION MODEL FORECAST EVALUATION

Year	P(delay$\leq$ 60min) =91.8%	P(delay$>$ $120min$) =2.78%
2010	90.9%	3.2%
2011	93.2%	2.2%
2012	91.3%	2.9%
2010~2012	92.3%	2.6%

Table 6.1A shows the validity of our model. Here we test the probability that a flight will be delayed within 60 minutes and probability it will be delayed more than 120 minutes. The actual probability is 91.8% and 2.78% respectively. We use our model to generate predictions on holdout set. 2011 has the minimum delay probability compared to other years.

2010 is the worst of flight delays and 2012 is in the middle. We can see when we evaluate the general performance of 2010~2012, the probabilities are around the mean value of sum of

2010, 2011 and 2012. Hence, our predictions have high accuracy which are all around actual probabilities.

6.2 Evaluation of Improved Real Time Arrival Prediction Model

In order to evaluate the performance of the whole arrival prediction model, here we will take Flight No.59 as an example first to compare ten days arrival predictions and actual arrival delays. We use training data starts from December 01, 2013 to February 18, 2014; and we will do predictions for the following 10 days which starts from February 19, 2014 to February 28, 2014. We will produce model delay values first, and then generate weather delay values by using real weather data for those 10 days. The final step is to add model delay value and weather delay value together to get one total prediction. After the example, we will give prediction results for all flights in American Airline at San Francisco International Airport.

Flight No.59:

1) Model Delay Prediction

   Figure 6.2A depicts model delay of Flight No.59 clearly, in which the red line is the fitting function generated by smooth.spline function in R and the blue line is our predictions for model delay for the following 10 days.

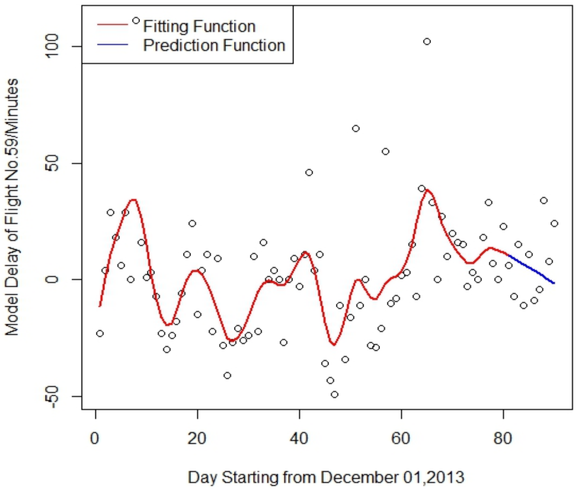


FIGURE 6.2A: MODEL DELAY OF FLIGHT NO.59

The predictions for the following 10 days:

```
[1] 10.4981002   9.1702696   7.8424391   6.5146085   5.1867779   3.8589473
[7]   2.5311168   1.2032862 -0.1245444 -1.4523750
```

2) Weather Delay Prediction

After fetching weather data consist of wind speed, visibility and sky conditions, we apply them to the formula we have created in the previous chapter. Then we get weather delay values for the following 10 days:

```
[1]   10.3508233   -23.0026533   -23.0026533   -23.0026533     0.7875018   -19.4917610
[7]   34.8689798   17.8838808   19.6639522   27.7043425
```

3) Total Delay Prediction

By adding Model Delay and Weather Delay predictions together, we get:

```
[1]   20.84892 -13.83238 -15.16021 -16.48804     5.97428 -15.63281    37.40010
[8]   19.08717    19.53941    26.25197
```

4) Differences Between Actual Delay and Prediction Delay

Actual Delay Values for the following 10 days:

```
[1]    6   -7   15   -11    11    -9    -4   34    8   24
```

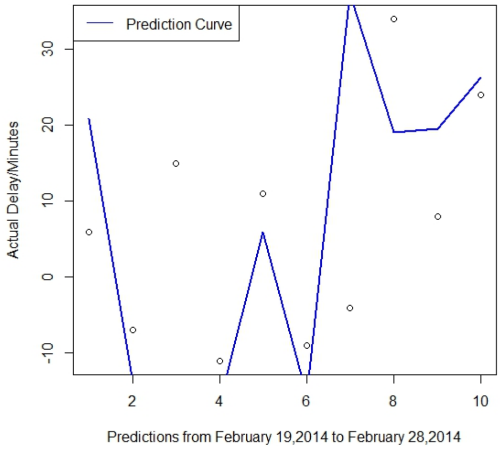


FIGURE 6.2 B ACTUAL ARRIVAL DELAY VS. PREDICTIONS

Here we can get the differences between actual delay and predictions as following:

```
[1] -14.848924    6.832384   30.160214    5.488045    5.025720    6.632814
[7] -41.400097   14.912833  -11.539408   -2.251968
```

Mean Value:

```
[1] -0.09883856
```

5) Results Analyses:

From the differences, we can see the smallest absolute value is only 2.251968 minutes and the largest is 41.400097 minutes. In general, the average mean difference is just 0.09883856 minutes which shows a good stability of our predictions.

Besides, if we set the absolute error range to 15 minutes, 90% of our predictions are within this error range, which means 90% possibility that we can do predictions with only 15 minutes absolute difference to actual delays.

Furthermore, our predictions here are for the following 10 days which depends on the accuracy of weather forecasts and this period model delay historical data; theoretically, our model could give a more higher accuracy when the prediction interval is shorter, usually we do predictions only for the next two days since the instability of weather forecasts.

Now, we will present all prediction performances in the following Table 6.2A: Prediction Performances of Improved Real Time Arrival Prediction Model from February 19, 2014 to February 28, 2014 of all flights of American Airline at San Francisco International Airport.

TABLE 6.2 A PREDICTION PERFORMANCES OF IMPROVED REAL TIME ARRIVAL PREDICTION MODEL

Flight No.	Prediction Accuracy(≤15 Mins)	Prediction Accuracy(≤30 Mins)
59	*80%*	*80%*
69	*60%*	*90%*
85	*70%*	*100%*
108	*70%*	*80%*
122	*90%*	*90%*
167	*70%*	*90%*
177	*90%*	*100%*
179	*80%*	*100%*
193	*80%*	*80%*
197	*90%*	*90%*
203	*80%*	*80%*
219	*90%*	*90%*
275	*90%*	*90%*
323	*80%*	*80%*
329	*80%*	*90%*
399	*70%*	*90%*
1105	*90%*	*100%*
1399	*90%*	*90%*
1427	*80%*	*80%*
1521	*80%*	*80%*
1524	*90%*	*90%*
1536	*80%*	*100%*
1585	*70%*	*90%*
1658	*80%*	*90%*
1677	*90%*	*90%*
2245	*80%*	*80%*
2356	*90%*	*90%*
2411	*90%*	*100%*
2455	*80%*	*100%*
2456	*80%*	*90%*
2457	*70%*	*100%*
2465	*90%*	*90%*
Mean	***81.25%***	***90%***

In sum, Table 6.2 A has presented a good proof that our model has an 81.25% accuracy if the absolute error range is 15 minutes and even a 90% accuracy when the absolute error range is 30 minutes. These data has proved the stability and tolerance of our model.

CHAPTER 7 WEB APPLICATION IMPLEMENTATION & PERFORMANCES

In this thesis, as stated in previous chapter, not only theory implementations have been illustrated, but also a web application with arrival delay prediction function will be shown. Hence, in this chapter, we will present a detailed description of the web system. It will explain the purpose and features of the system, the interfaces of the system, software specifications, web page framework and prediction function implementations of our web page.

7.1 Scope of Project

This software system will be a web application for predictions of flight arrival delays. More specifically, this system is designed to allow the user to search for flights satisfying the user criteria which allows searching flights of American Airline at San Francisco International Airport at present. The software enables users to track a flight status, whether it has arrived or how many minutes it will be delayed according to the model we have built during this thesis. Besides, other basic information of the flight will be returned together in a fixed format. Hence, this web page not only facilitates a better understanding of our model, but also provides a good experience for users to explore. Furthermore, this web page can be practically developed for predicting arrival delays of all flights for all airports.

7.2 Web Application Design Overview

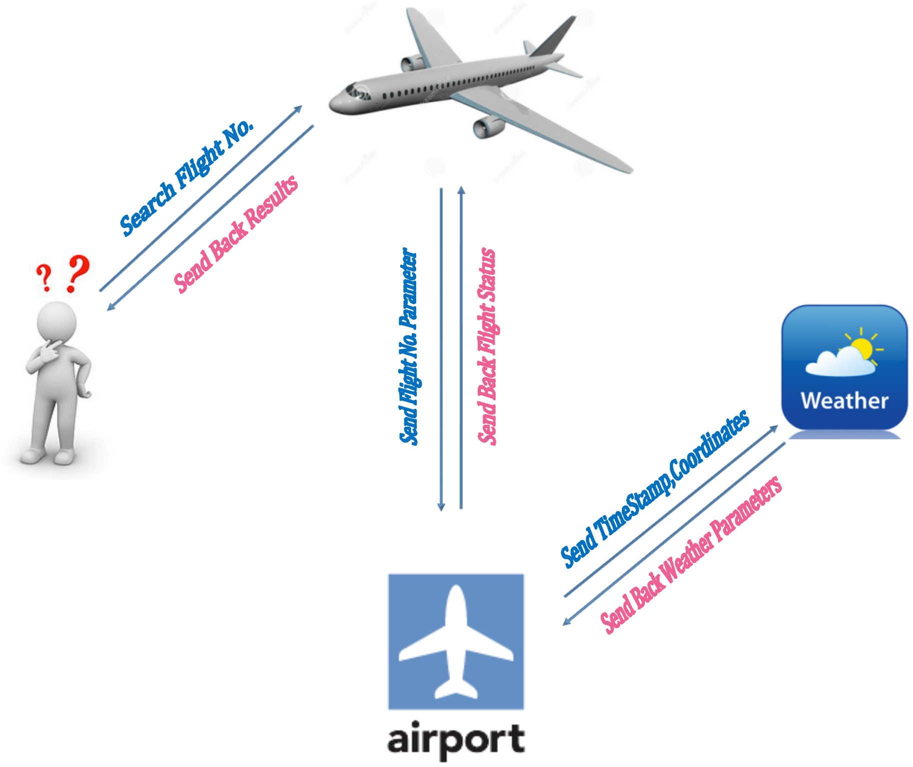

Search Flight No.
Send Back Results
Send Flight No. Parameter
Send Back Flight Status
Weather
Send TimeStamp,Coordinates
Send Back Weather Parameters
airport

FIGURE 7.1A: WEB APPLICATION DESIGN OVERVIEW

There are four classes in our web application. Figure 7.1A visually illustrates how these four classes interact with each other and also demonstrates how different functions implemented. With the illustration of this figure, now we will explain more about how the whole system works with the following details.

Firstly, the web application enables users to search daily American Airline flights at San Francisco International Airport. To obtain the flight status users want, users only need to type in flight number among American Airline.

Secondly, the flight number will be used as the pass parameter to a **Flight Schedule API** which is provided by flightstats.com website. The Schedule API can provide up-to-date flight schedule information including scheduled departure time, scheduled arrival time, departure terminal, arrival terminal, actual departure time and actual arrival time if the flight has departed or has arrived.

Thirdly, after obtaining scheduled arrival time, our system will detect whether this flight has arrived or not. If the flight has arrived, then all basic information from Flight Schedule API will be sent as a result to front end to give users searching results.

However, if the flight has not arrived yet, then our system will launch an arrival delay prediction for this flight by using models we have built in this thesis. As we have introduced in our model, in order to do the prediction, we need parameters/data for both Model Delay and Weather Delay Models. **As to Model Delay prediction**, the system will fetch the latest historical arrival model delay data of the flight and loaded into our database. After the storage of the data, the smooth.spline function will be applied to produce the next Model Delay value and also will be stored in the database.

At the same time, another part is Weather Delay prediction; here we will also utilize a **Weather API** provided by the same website to acquire all weather parameters we need. With the purpose of acquiring accurate weather information, the scheduled arrival time we fetched from Flight Schedule API will be used as the pass parameter to get the forecast weather information. After obtaining all weather data including wind speed, visibility and sky conditions, and the system will apply the multiple linear regression function we have produced to generate the Weather Delay value.

Lastly, Model Delay value and Weather Delay value will be summed together to form the final arrival prediction value of the flight the user has searched. And this result will be returned back to the user with other flight information from Flight Schedule API.

7.3 Software Specification

7.3.1 Programming Language

This web application is mainly implemented through **Java with Maven**, which is a build automation tool and we also use **RCaller**, which is a software library to call **R language from Java**. We have the following reasons why choose them.

A. The Advantages of **Java**[33]:

1) Java is designed to be easily used, compiled, debugged and run.

2) Java is object-oriented which allows users to reuse codes and do maintenances.

3) Java is also platform-independent, which can be regarded as the most significant advantages. We can move Java Projects easily from one system to another. This ability makes Java succeeds.

4) Java is distributed which is designed to make distributed computing simple with the networking capability that is inherently integrated into it. Writing network programs in Java is like sending and receiving data to and from a file.

5) Java is secure, which is shown in each development like compiler, interpreter, and runtime environment. It considers security as part of its design.

6) Java is robust also, this is reflected by java compilers are able to detect many problems that would first show up during execution time in other languages.

7) Java is even multithreaded, which has the capability for a program to perform several tasks simultaneously within one program. In other languages, operating system-specific procedures have to be called in order to enable multithreading. However, in Java, multithreaded programming has been smoothly integrated into it.

B. The Advantages of **Maven**[34]:

1) Simple project setup that follows best practices - get a new project or module started in seconds.

2) Consistent usage across all projects means no ramp up time for new developers coming onto a project.

3) Superior dependency management including automatic updating, dependency closures (also known as transitive dependencies).

4) Able to easily work with multiple projects at the same time.

5) A large and growing repository of libraries and metadata to use out of the box, and arrangements in place with the largest Open Source projects for real-time availability of their latest releases.

6) Extensible, with the ability to easily write plugins in Java or scripting languages.

7) Instant access to new features with little or no extra configuration.

8) Ant tasks for dependency management and deployment outside of Maven.

9) Model based builds: Maven is able to build any number of projects into predefined output types such as a JAR, WAR, or distribution based on metadata about the project, without the need to do any scripting in most cases.

10) Coherent site of project information: Using the same metadata as for the build process, Maven is able to generate a web site or PDF including any documentation you care to add, and adds to that standard reports about the state of development of the project. Examples of this information can be seen at the bottom of the left-hand navigation of this site under the "Project Information" and "Project Reports" submenus.

11) Release management and distribution publication: Without much additional configuration, Maven will integrate with your source control system such as CVS and manage the release of a project based on a certain tag. It can also publish this to a distribution location for use by other projects. Maven is able to publish individual outputs such as a JAR, an archive including other dependencies and documentation, or as a source distribution.

12) Dependency management: Maven encourages the use of a central repository of JARs and other dependencies. Maven comes with a mechanism that your project's clients can use to download any JARs required for building your project from a central JAR repository much like Perl's CPAN. This allows users of Maven to reuse JARs across projects and encourages communication between projects to ensure that backward compatibility issues are dealt with.

C. The advantage of RCaller:

RCaller is a software library for Calling R from Java, which is easy to use and implemented. There are many software libraries that communicate R with other languages. One of them is called JRI, which allows to run R inside Java applications as a single thread. Basically it loads R dynamic library into Java and provides a Java API to R functionality. It supports both simple calls to R functions and a full running REPL. But the weak point of JRI is the complex environment setting for a fresh hand. It usually costs long time to install and to complete the whole configurations. In contrast, RCaller is a software library which comes into prominence with its simplicity. The only thing need to do is add RCaller to the corresponding Java Library.

In sum, the above content shows why we utilize these programming languages or software. The following Figure 7.3A can show the relationship among them more explicitly.

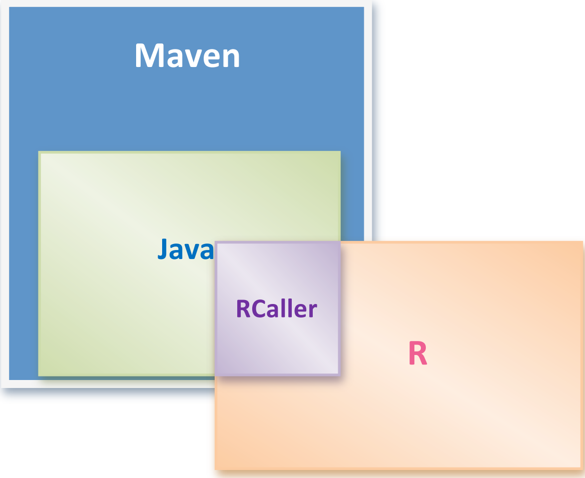


FIGURE 7.3A: RELATIONSHIP AMONG MAVEN, JAVA, RCALLER AND R

7.3.2 Web Server

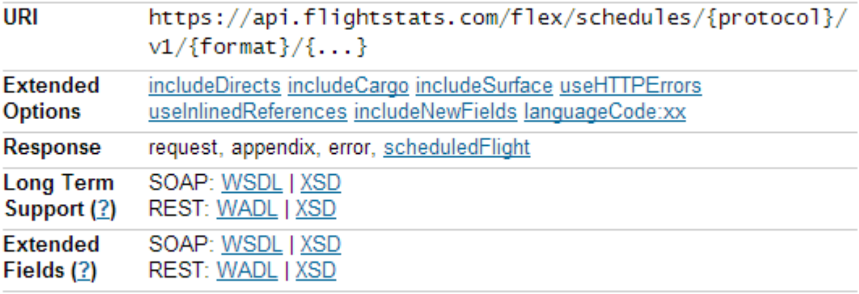
We use Apache Tomcat as our web server. It is an open source software implementation of the Java Servlet and JavaServer Pages technologies. Besides, Apache Tomcat is developed in an open and participatory environment, which is intended to be a collaboration of the best-of-breed developers from around the world. It also powers numerous large-scale, mission-critical web applications across a diverse range of industries and organizations.

7.4 Application Programming Interface

After introducing the above important technologies and software we have employed in our web page implementation, now we will present two important Application Programming Interfaces (APIs) in our project. As we have mentioned in the previous content, the two APIs we have adopted from flightstats.com website are Schedules API and Weather API.

7.4.1 Schedules API [35]

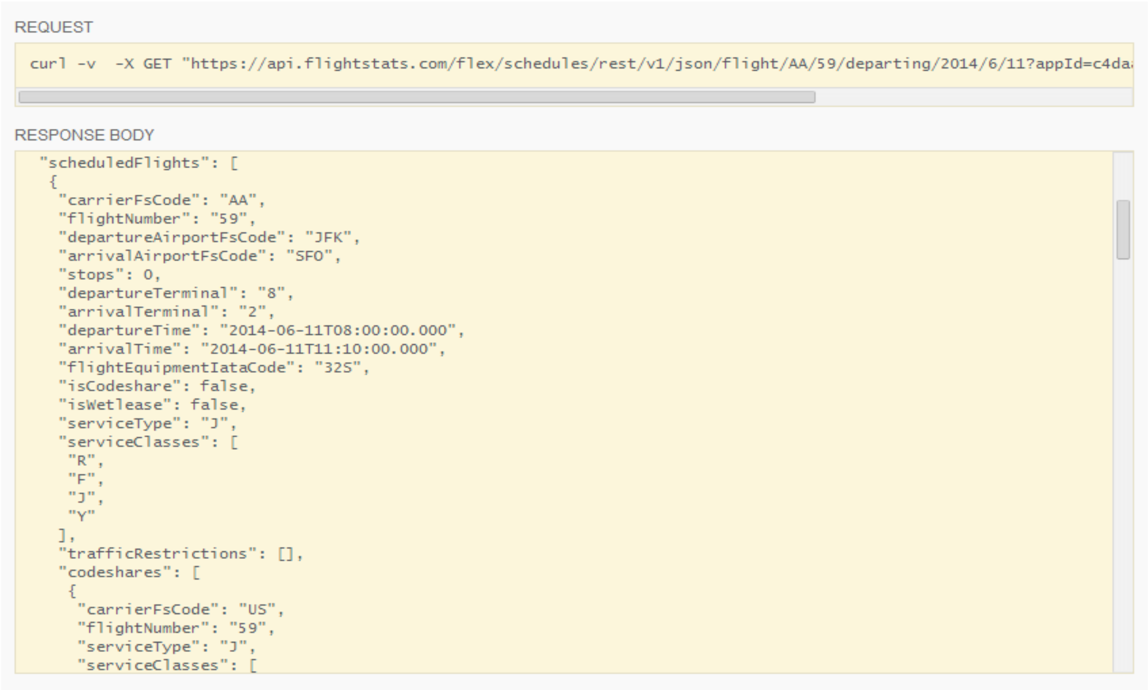
The Schedules API offers information on future scheduled flights. The following Table 7.4A: Schedules API presents basic information of this API.



URI	`https://api.flightstats.com/flex/schedules/{protocol}/v1/{format}/{...}`
Extended Options	includeDirects includeCargo includeSurface useHTTPErrors useInlinedReferences includeNewFields languageCode:xx
Response	request, appendix, error, scheduledFlight
Long Term Support (?)	SOAP: WSDL | XSD REST: WADL | XSD
Extended Fields (?)	SOAP: WSDL | XSD REST: WADL | XSD

Table 7.4A: Schedules API

Here is an example about the response of Schedules API when we search Flight No.59 on June, 11, 2014:



REQUEST

```
curl -v  -X GET "https://api.flightstats.com/flex/schedules/rest/v1/json/flight/AA/59/departing/2014/6/11?appId=c4da
```

RESPONSE BODY

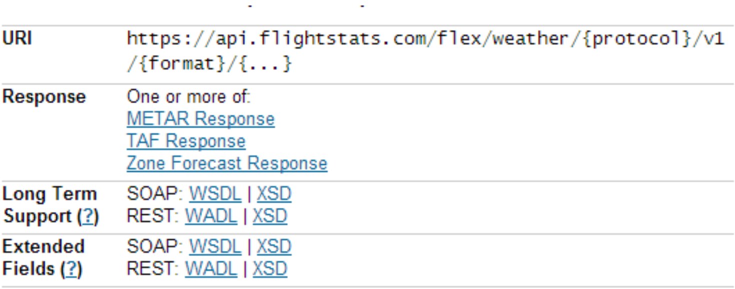
```
  "scheduledFlights": [
   {
    "carrierFsCode": "AA",
    "flightNumber": "59",
    "departureAirportFsCode": "JFK",
    "arrivalAirportFsCode": "SFO",
    "stops": 0,
    "departureTerminal": "8",
    "arrivalTerminal": "2",
    "departureTime": "2014-06-11T08:00:00.000",
    "arrivalTime": "2014-06-11T11:10:00.000",
    "flightEquipmentIataCode": "32S",
    "isCodeshare": false,
    "isWetlease": false,
    "serviceType": "J",
    "serviceClasses": [
     "R",
     "F",
     "J",
     "Y"
    ],
    "trafficRestrictions": [],
    "codeshares": [
     {
      "carrierFsCode": "US",
      "flightNumber": "59",
      "serviceType": "J",
      "serviceClasses": [
```

FIGURE 7.4A: RESPONSE EXAMPLE OF SCHEDULES API

From the response, we can observe that all detailed information about Flight No.59 is presented in the content. Among these data, we will extract "arrivalTime" value and send as the pass parameter when we request Weather API.

7.4.2 Weather API [36]

TABLE 7.4B: WEATHER API



URI	https://api.flightstats.com/flex/weather/{protocol}/v1 /{format}/{...}
Response	One or more of: METAR Response TAF Response Zone Forecast Response
Long Term Support (?)	SOAP: WSDL | XSD REST: WADL | XSD
Extended Fields (?)	SOAP: WSDL | XSD REST: WADL | XSD

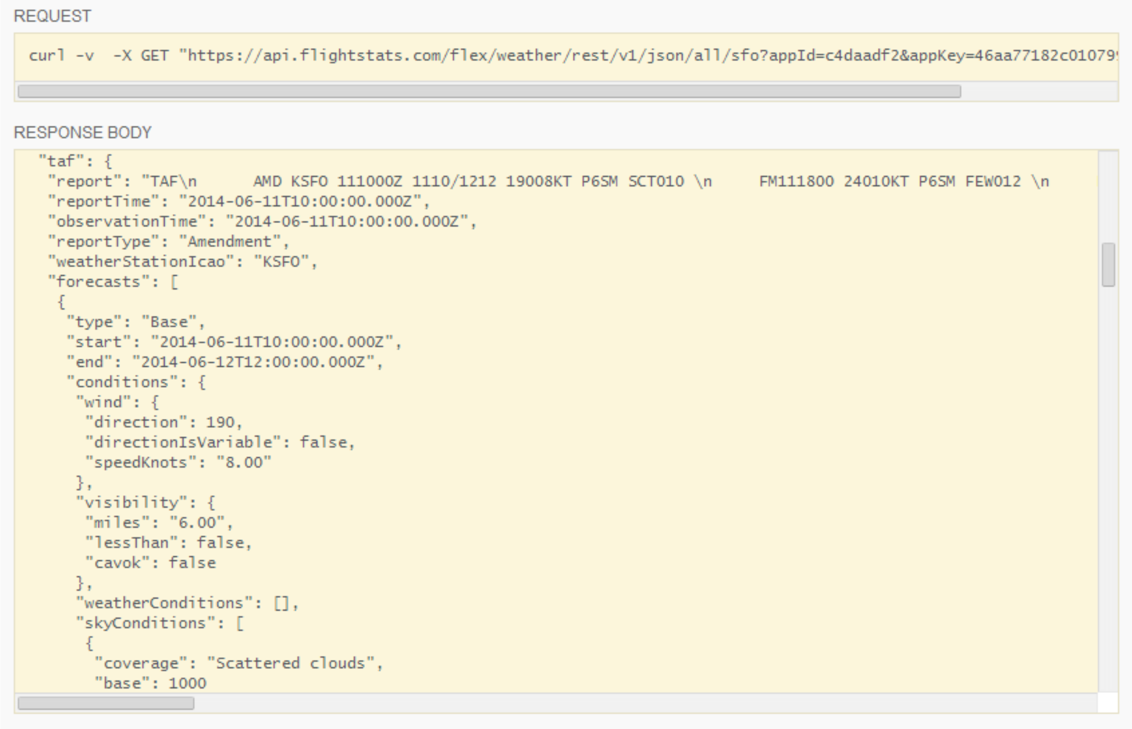
Table 7.4B: Weather API shows basic information and the The Weather API includes METAR, TAF (Terminal Aerodrome/Area Forecast), and Zone Forecasts. Weather information is currently available only for locations in the United States.

METAR reports provide up-to-date information on current weather conditions at an airport; we enrich these reports with tag annotations identifying prevailing conditions and notable hazards that may impact aviation.
Our two forecast products, TAF and Zone Forecasts, are complementary in scope.

TAF provides a detailed forecast for the immediate vicinity of an airport, generally covering a 9 to 12 hour window (sometimes greater).

Zone Forecasts provide longer-term and more geographically broad outlook via day-by-day forecasts generally stretching up to about a week in the future.

Here is also attached an example about the response of Weather API when we search Flight No.59 on June, 11, 2014:



```
curl -v  -X GET "https://api.flightstats.com/flex/weather/rest/v1/json/all/sfo?appId=c4daadf2&appKey=46aa77182c01079
```

```
  "taf": {
   "report": "TAF\n      AMD KSFO 111000Z 1110/1212 19008KT P6SM SCT010 \n     FM111800 24010KT P6SM FEW012 \n
   "reportTime": "2014-06-11T10:00:00.000Z",
   "observationTime": "2014-06-11T10:00:00.000Z",
   "reportType": "Amendment",
   "weatherStationIcao": "KSFO",
   "forecasts": [
    {
     "type": "Base",
     "start": "2014-06-11T10:00:00.000Z",
     "end": "2014-06-12T12:00:00.000Z",
     "conditions": {
      "wind": {
       "direction": 190,
       "directionIsVariable": false,
       "speedKnots": "8.00"
      },
      "visibility": {
       "miles": "6.00",
       "lessThan": false,
       "cavok": false
      },
      "weatherConditions": [],
      "skyConditions": [
       {
        "coverage": "Scattered clouds",
        "base": 1000
```

FIGURE 7.4B: RESPONSE EXAMPLE OF WEATHER API

Since Flight No.59 will arrive at SFO around 11:10, we just fetch the weather information which has the closest timestamp to 11:10. In the response, we get “forecasts” starts from “2014-06-11T10:00:00.000Z” to “2014-06-12T12:00:00.000Z”. There are three parameters we need which are wind speed, visibility and sky conditions. Wind speed here is reflected by speedKnots in this response, which is “8.00”. Visibility is “6.00” miles and sky conditions are “Scattered clouds”. After getting these parameters, we will convert these Strings to real numbers we have defined in Table 5.2E and Table 5.2F. The last step is to apply these numbers to our multiple linear regression function to produce the weather delay value.

7.5 Main Software Versions

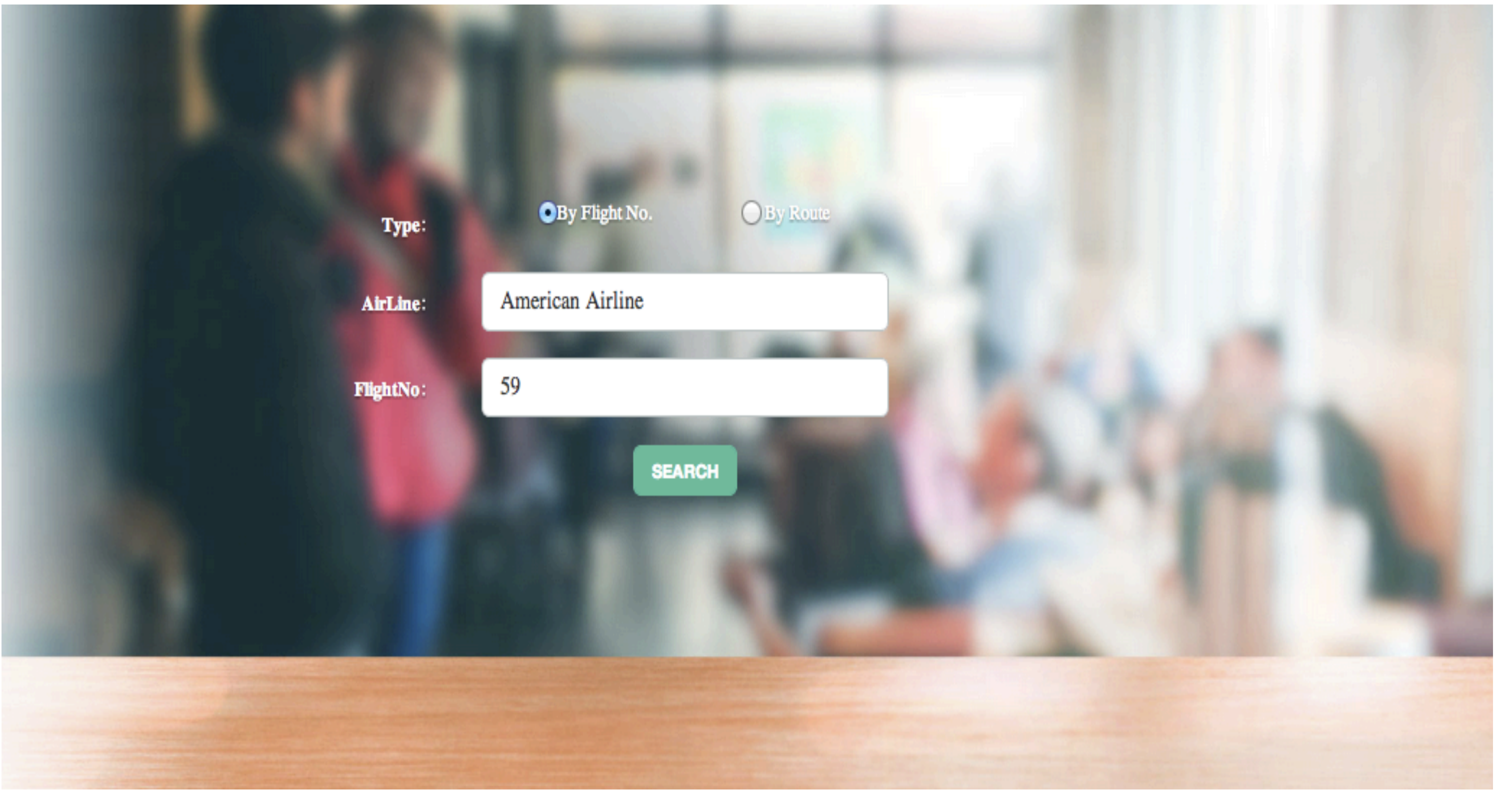
Eclipse IDE for Java EE Developers Mac OS X(Cocoa 32)

Tomcat 7

RCaller 2.2

7.6 Interface Design

7.6.1 Main Searching Page



Type: ● By Flight No. ○ By Route

AirLine: American Airline

FlightNo: 59

FIGURE 7.6A: MAIN SEARCHING PAGE

- Clear & Easy-use Searching Function Design

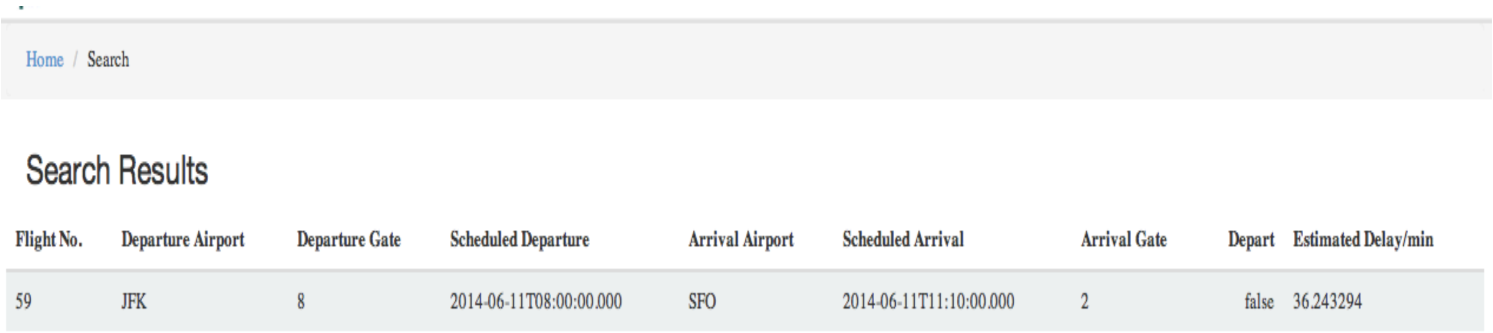
From Figure 7.6A, we can see our beautiful home page which has a very clear searching module. This character makes users very easy to know how to search the flight they want to track. There are two types of searching methods. First one is searching by Flight Number, which is available now at our website. The other option is searching by route which will be developed in the future. The Airline text box is designed from all possible airlines. Now we will use American Airline as an illustration of the model we have built during this thesis. The last text box is for writing a flight number of American Airline. Figure 7.6 takes Flight No.59 as an example here. After inputting these parameters, we can click on SEARCH button to see the arrival delay prediction for Flight No.59.

● Comfortable Background & Elegant Design Style

Besides the above advantages we have shown, our page is using a comfortable background picture, which simulates the real scene of passengers waiting at the airport for their flights. Meanwhile the whole layout and color assortment of the main page reflects the elegance of our arrival delay prediction web page. Users will have an immersive experience by using this web page.

7.6.2 Searching Results Page

FIGURE 7.6B: SEARCHING RESULTS PAGE



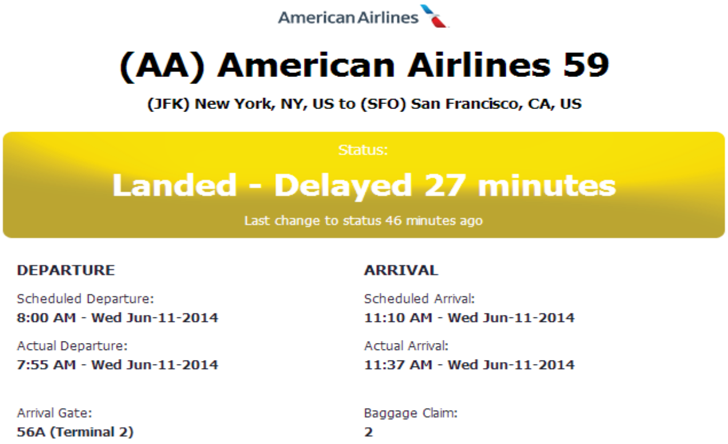
Home / Search

Search Results

Flight No.	Departure Airport	Departure Gate	Scheduled Departure	Arrival Airport	Scheduled Arrival	Arrival Gate	Depart	Estimated Delay/min
59	JFK	8	2014-06-11T08:00:00.000	SFO	2014-06-11T11:10:00.000	2	false	36.243294

As shown in Figure 7.6B, the search results returned information consists of Flight No., Departure Airport, Departure Gate/Terminal, Scheduled Departure time, Arrival Airport, Scheduled Arrival time, Arrival Gate/Terminal, Depart Boolean value which means the flight has departed or not and Estimated Delay time. Here we give a specific estimated arrival delay prediction instead of range possibilities, which is usually provided by other flight prediction website. Since a specific delay value is more directly for users and can be more accurate than range predictions.

Besides, the following Figure 7.6C: Actual Arrival Delay of Flight No.59 [] shows the actual delay is 27 minutes after Flight No.59 has landed. Compared to our prediction 36 minutes in Figure 7.6 B, we can see our estimated arrival delay prediction has a good accuracy. This example is one of the excellent proofs of our prediction model.



American Airlines

(AA) American Airlines 59

(JFK) New York, NY, US to (SFO) San Francisco, CA, US

Status:

Landed - Delayed 27 minutes

Last change to status 46 minutes ago

DEPARTURE	ARRIVAL
Scheduled Departure: 8:00 AM - Wed Jun-11-2014	Scheduled Arrival: 11:10 AM - Wed Jun-11-2014
Actual Departure: 7:55 AM - Wed Jun-11-2014	Actual Arrival: 11:37 AM - Wed Jun-11-2014
Arrival Gate: 56A (Terminal 2)	Baggage Claim: 2

FIGURE 7.6C: ACTUAL ARRIVAL DELAY OF FLIGHT NO.59

CHAPTER 8 CONCLUSIONS & FUTURE WORK

As a conclusion, both the two models we have created for this thesis show excellent performances and tolerances. The first model is for long term prediction and the second one is for short term real time prediction by using big real time data. Hence, they can be utilized by passengers to obtain flight status and also can be applied for customers to do risk management in different fields.

As for future work, we will improve and extended in two sides. One side is to seek a more accurate model to increase the accuracy of our predictions. As stated in introduction part, Bayesian theory is a good direction to further develop. Even though the priori probability is not easy to acquire, we can utilize some mathematical theories and years of historical data to generate the priori probability first. This will be conducive if we can combine the priori probability with real data. It will not only yield to a finest prediction model but also avoid over fit problem of data. The other side we can do is to develop our web page with more functions. Now the web page is mainly for American Airline at San Francisco International Airport, but it has the potentiality to extend for all airlines and all airports. By combining these two sides, we believe our model and web page will contribute more in flight delay risk management area.

REFERENCES

[1] E.R. Mueller and G.B. Chatterji. “Analysis of aircraft arrival and de-parture delay characteristics”. In: Proceedings of the AIAA Aircraft Technology, Integration, and Operations (ATIO) Conference, Los Angeles, CA. 2002.

[2] Lu Zonglei, Wang Jiandong, and Zheng Guansheng. “A new method to alarm large scale of flights delay based on machine learning”. In: Knowledge Acquisition and Modeling, 2008. KAM ’08. International Symposium on. 2008, pp. 589–592.

[3] Yu-jie Liu et al., Flight delay propagation research based on Bayesian network, Computer Engineering and Applications (In Chinese), 44, No.17 (2008), 242-245.

[4] Yu-Jie Liu, Wei-Dong Cao, and Song Ma, Estimation of arrival flight delay and delay propagation in a busy hub-airport, The Proceedings of the IEEE ICNC’08: 2008 Fourth International Conference on Natural Computation, 4 (2008), 500-505.

[5] Zhang Lianwen and Guo Haipeng, Introduction to Bayesian Networks, Science Publication, China (2006), 31-41.

[6] Y. Tu, M.O. Ball, and W.S. Jank. “Estimating flight departure delay distributions – a statistical approach with long-term trend and short-term pattern”. In: Journal of the American Statistical Association 103.481 (2008), pp. 112–125.

[7] Research and Innovative Technology Administration.(March 2014). San Francisco International Airport Snapshot. Retrieved from http://www.transtats.bts.gov/airports.asp?pn=1

[8] Research and Innovative Technology Administration.(March 2014). Denvor International Airport Snapshot. Retrieved from http://www.transtats.bts.gov/airports.asp?pn=1

[9,10] Research and Innovative Technology Administration.(March 2014). Carrier Snapshot. Retrieved from http://www.transtats.bts.gov/carriers.asp

[11] Research and Innovative Technology Administration.(Dec 2013). Understanding the Reporting of Causes of Flight Delays and Cancellations. Retrieved from http://www.rita.dot.gov/bts/help/aviation/html/understanding.html

[12] Research and Innovative Technology Administration.(Dec 2013). Weather's Share of Total Delay Minutes. Retrieved from http://www.rita.dot.gov/bts/help/aviation/html/understanding.html

[13] Carl de Boor. A Practical Guide to Splines. Springer, New York, 1978.

[14] Mills, Terence C. (1990) Time Series Techniques for Economists. Cambridge University Press.

[15] Petrevska, Biljana. 2012. Forecasting international tourism demand: The evidence of Macedonia. *UTMS Journal of Economics* 3 (1): 45–55.

[16] 2 Dead After Boeing 777 Crashes on Landing at San Francisco International Airport. Retrieved from http://newsfeed.time.com/2013/07/06/boeing-777-crash-lands-at-san-francisco-international-airport/

[17] K. Takezawa. Introduction to nonparametric Regression. Hoboken, New Jersy,2006

[18] Charles,Zaiontz. 2014. Real Statistics Using Excel.Retrieved from http://www.real-statistics.com/multiple-regression/multiple-regression-analysis/

[19] Jolliffe, Ian T. 1982. "A Note on the Use of Principal Components in Regression". *Journal of the Royal Statistical Society, Series C* **31** (3): 300–303.

[20] Bates, D. M. & D. G. Watts. 1988. Nonlinear Regression Analysis and Its Applications. New York: Wiley.

[21] Gallant, A. R. 1975. “Nonlinear Regression.” The American Statistician 29:73—81.

[22] K. Takezawa. Introduction to nonparametric Regression. Hoboken, New Jersy,2006

[23] Rachelle,Oblack. 2014.Visibility. Retrieved from http://weather.about.com/od/v/g/visibility.html

[24,27,28] Experimental Aircraft Info. 2014. Retrieved from http://www.experimentalaircraft.info/wx/weather-visibility.php

[25] "782 – Aerodrome reports and forecasts: A user's handbook to the codes". *World Meteorological Organization*. Retrieved 2009-09-23.

[26] "Chapter 7". Aeronautical Information *Manual*. Retrieved 2007-12-01.

[29] Jayanta,Basak. Weather Data Mining Using Independent Component Analysis. The Journal of Machine Learning Research. Volume 5, 12/1/2004 Pages 239-253

[30] 2 Dead After Boeing 777 Crashes on Landing at San Francisco International Airport. Retrieved from http://newsfeed.time.com/2013/07/06/boeing-777-crash-lands-at-san-francisco-international-airport/

[31] K. Takezawa. Introduction to nonparametric Regression. Hoboken, New Jersy,2006

[32] Mark, Harmon. March,2010. Regression - How To Quickly Read the Output of Excel's Regression. Retrieved from http://blog.excelmasterseries.com/2010/03/how-to-quickly-read-output-of-excels.html

[33] Miloš,Bielik.2011.Seven Advantages of Java. Retrieved from http://bielik.blogspot.no/2011/05/seven-advantages-of-java.html

[34] Apache Maven Project.2014. Retrieved from http://maven.apache.ort/maven-features.html

[35] FlightStats Devlop Center. 2014. Schedule API. Retrieved from https://developer.flightstats.com/api-docs/scheduledFlights/v1

[36] FlightStats Devlop Center. 2014. Weather API. Retrieved from https://developer.flightstats.com/api-docs/weather/v1