Delay Characterization and Prediction in Major U.S. Airline Networks

by

Zebulon James Hanley

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

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author .....
Sloan School of Management
May 8, 2015

Certified by .....
Hamsa Balakrishnan
Associate Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by .....
Dimitris Bertsimas
Boeing Leaders for Global Operations Professor of Management
Professor of Operations Research
Co-director, Operations Research Center

Delay Characterization and Prediction in Major U.S. Airline Networks

by

Zebulon James Hanley

Submitted to the Sloan School of Management
on May 8, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

This thesis expands on models that predict delays within the National Airspace System (NAS) in the United States. We propose a new method to predict the expected behavior of the NAS throughout the course of an entire day after only a few flying hours have elapsed. We do so by using k-means clustering to classify the daily NAS behavior into a small set of most commonly seen snapshots. We then use random forests to map the delay behavior experienced early in a day to the most similar NAS snapshot, from which we make our type-of-day prediction for the NAS. By noon EST, we are able to predict the NAS type-of-day with 85% accuracy. We then incorporate these NAS type-of-day predictions into previously proposed models to predict the delay on specific origin-destination (OD) pairs within the U.S. at a certain number of hours into the future. The predictions use local delay variables, such as the current delay on specific OD pairs and airports, as well network-level variables such as the NAS type-of-day. These OD pair delay prediction models use random forests to make classification and regression predictions. The effects of changes in classification threshold, prediction horizon, NAS type-of-day inclusion, and using wheel off/on, actual, and scheduled gate departure and arrival times are studied. Lastly, we explore how the delay behavior of the NAS has changed over the last ten years and how well the models perform on new data.

Thesis Supervisor: Hamsa Balakrishnan
Title: Associate Professor of Aeronautics and Astronautics

Acknowledgments

I would like to thank my advisor, Professor Hamsa Balakrishnan, for her guidance and insightful ideas. Without her as an advisor this work would not have been possible. I would also like to thank the tremendous students at the MIT Operations Research Center who proved to be invaluable academic resources and amazing friends. I would also like to thank the members of my family for their never-ending love and support and for teaching me that with diligence anything is possible.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The National Airspace System (NAS) in the United States is a fundamental component of the economy, national security, and the American way of life. Analysis of the U.S. airline network is necessary to better understand how to ensure its efficient operation. In 2014, 21% of U.S. flights were delayed by greater than 15 minutes, and over 2.2% of flights were cancelled[8]. These domestic flight delays produce significant economic losses. The FAA Consortium in Aviation Operations Research estimates that in 2007 domestic flight delays cost the U.S. economy $31.2 billion [1].

Air travel has always been influenced by weather, but network-related delays and congestion are playing an increasingly significant role in the efficiency of the NAS today. Airlines continually strive for operational efficiency by increasing resource utilization; simultaneously, demand for air travel threatens to exceed capacity constraints in the network. Combined, these conditions create a network that is sensitive to delays. A drive to increase resource utilization decreases inter-flight buffer times, and near-100% utilization provides for little dynamic flexibility [11]. Thus, delays are increasingly being propagated through the network over time and distance.

This thesis attempts to explain and model some of these network effects. Its primary goal is to improve upon delay prediction models that employ network-based variables. In particular, we are interested in predicting the departure delays of particular origin-destination (OD) pairs by considering past, current, and predicted future delay states of different network elements. Because of the interdependency of elements

within the U.S. airline network, we believe that the current status and behavior of certain network elements can yield meaningful predictions about delays in the short-term future. Because of the emphasis on network-related effects, we do not attempt to properly model microscopic delay events, such as those caused by a single aircraft mechanical problem, but rather we concentrate our efforts of modeling macroscopic delay behaviors that manifest themselves on a larger regional or national level. Thus, we do not anticipate our analysis to be particularly useful in predicting individual flight delays, but we expect it to be beneficial when looking at aggregated delay levels. Nevertheless, we perform our analysis and evaluate prediction performance on data that includes all sources of delay. While some data sources (such as the Bureau of Transportation Transtats database [7]) attempt to identify the cause for individual flight delays, we find such identification to be incomplete, inaccurate, or too ambiguous to be considered.

Aircraft delay prediction has been continually studied by the research community. Various prediction models have been proposed. Weather is a primary factor in airline delays and is commonly studied, such as in [5] and [4]. [2] uses an agent-based data-driven model to reproduce delay propagation during an severe weather event. Emphasis has also been placed on aircraft rotations and the connectivity of passengers and crews as a significant source of delay spreading [3]. Stochastic models using Bayesian Networks have been employed [6, 13]. Both [6] and [13] adopt a similar approach of utilizing local, microscopic delay variables while employing Bayesian Networks for prediction. Recently, stochastic methods have also been employed in a dynamic queue engine to model delay propagation [9]. Delay propagation has also been modeled using survival models for a single Taiwanese airline [12]. This thesis draws heavily upon work performed in the masters thesis *Characterization and Prediction of Air Traffic Delays* submitted by Juan Jose Rebollo [10], which uses clustering techniques to create variables indicative of the state of the NAS at a given time. In contrast to other works, [10] investigates the ability of a single NAS delay variable to characterize the behavior of the U.S. airline network and to predict delays. Many of the preprocessing steps taken in this thesis parallel those conducted in [10].

In this thesis we create delay prediction models to predict the departure delays on specific OD pairs. To do so, we explore the behavior of the NAS and the feasibility of modeling its status using a single variable for its hourly and daily behavior. We expand upon [10] by updating the NAS state and NAS type of day clustering on newer data while also including a discussion about how this method returns different results for different years. We explore how the use of different time definitions of events (e.g. actual gate departure versus scheduled gate departure) affect prediction models. Additionally, we create models to predict the NAS state in a given hour in the future and to predict the overall behavior of the NAS throughout the course of a day. We incorporate these variables, in additional to temporal variables (such as the time of day, day of week, and season) and local delay variables (such as the expected delay on a specific OD pair) to conduct the delay predictions. Similar to [10], we identify which variables are important in delay prediction, and we choose a subset of all possible variables in order to simply the delay prediction models. Both classification and regression models are investigated.

1.1 Thesis Organization

This thesis is organized as follows: Chapter 2 describes the dataset used in this research and initial analysis and preprocessing performed on the data. Chapter 3 introduces a variable to describe the overall state of the NAS in a given hour and characterizes typical behavior in the U.S. airline network. The goal of this chapter is to identify the main delay patterns in the NAS. Similar to Chapter 3, Chapter 4 introduces a variable to describe the overall state of the NAS in a given day (instead of a given hour) and characterizes typical behavior in the U.S. airline network. Chapter 4 also provides a model for predicting the daily behavior of the NAS after observing only a few elapsed ours in the day. The variables introduced in Chapters 2, 3, and 4 are used in Chapter 5 to create various models to predict the departure delay on a particular OD pair. Chapter 6 uses Monte Carlo cross-validation (MCCV) to investigate the performance of delay prediction models for the top 100 most-delayed

OD pairs. Chapter 7 presents various analysis designed to describe how the behavior of the NAS has changed over the past decade. Chapter 8 presents models to predict the NAS state variable defined in Chapter 3 for a given number of hours in the future. Finally, conclusions and next steps of this research are discussed in Chapter 9.

Chapter 2

Dataset Overview, Preprocessing, and Analysis of Temporal Variables

In this chapter we describe the data used in this thesis and the preprocessing steps conducted prior to model formulations. All data was obtained from the U.S. Bureau of Transportation Statistics Research and Innovation Technology Administration Transtats database. The BTS On-Time Performance database [7] logs details of individual flights reported by U.S. certified air carriers that account for at least one percent of domestic scheduled passenger revenues. We initially use two years of data from 2011 and 2012, which incorporates over 10 million flights and over 4,000 unique airports. Unless otherwise specified, in this thesis the data we use comprises of these two years of data. In Chapters 4, 6, and 7, we use also use data from other years for testing and for diagnosing the change in NAS behavior over the last decade. The following information is gathered for each flight:

- Flight Date
- Origin airport
- Destination airport
- Scheduled gate departure time
- Actual gate departure time
- Actual wheel-off departure time

- Departure delay (expressed as the difference between scheduled and actual gate departure time)
- Scheduled gate arrival time
- Actual gate arrival time
- Actual wheel-on arrival time
- Arrival delay (expressed as the difference between scheduled and actual gate arrival time)
- Cancellation (whether the flight was cancelled)
- Diverted (whether the flight was diverted mid-air)
- Tail number (unique tail number of the aircraft)

For the departure and arrival delay variables, any flights that were ahead of schedule (expressed as a negative delay) were zeroed. One possible future area of study is to re-do the analysis presented in this thesis with non-zeroed data, such that the models also predict if a flight runs ahead of schedule.

Flight times and dates in the Transtats database are expressed in local times, so we convert all the flight times and dates to be expressed in Eastern Standard Time (EST).

2.1 Network Simplification

Many of the preprocessing steps in this thesis are similar to those used in [10]. The first is network simplification. Because our analysis aims to find macroscopic network-based effects, we limit our analysis to OD pairs in the U.S. that demonstrate a significant influence on the rest of the network.

We define an OD pair as being directional, so a route from BOS to JFK is distinct from the route from JFK to BOS. We choose to limit our analysis to only include flights occurring on OD pairs that have at least 5 flights per day on average over the course of 2011-2012. Figure 2-1 displays the results of the simplification. It is evident that a small percent of OD pairs accounts for a large percent of total flight volume in

the United States. With our criteria of at least 5 flights per day on average, we retain over 60 percent of the original flights in the dataset, but we have reduced the network to the top 25 percent most frequented routes. This results in 158 airports, 1,107 OD pairs and over 8 million flights in the simplified dataset. We achieve this simplification without sacrificing a significant amount of meaningful data. We believe these most frequent routes capture the vast majority of NAS behavior and we therefore use this simplified network in the remainder of the project. Furthermore, these routes are relatively spread out across the United States, allowing for a good representation of the entire NAS. A U.S. map of the OD pairs in the simplified network is shown in Figure 2-2.

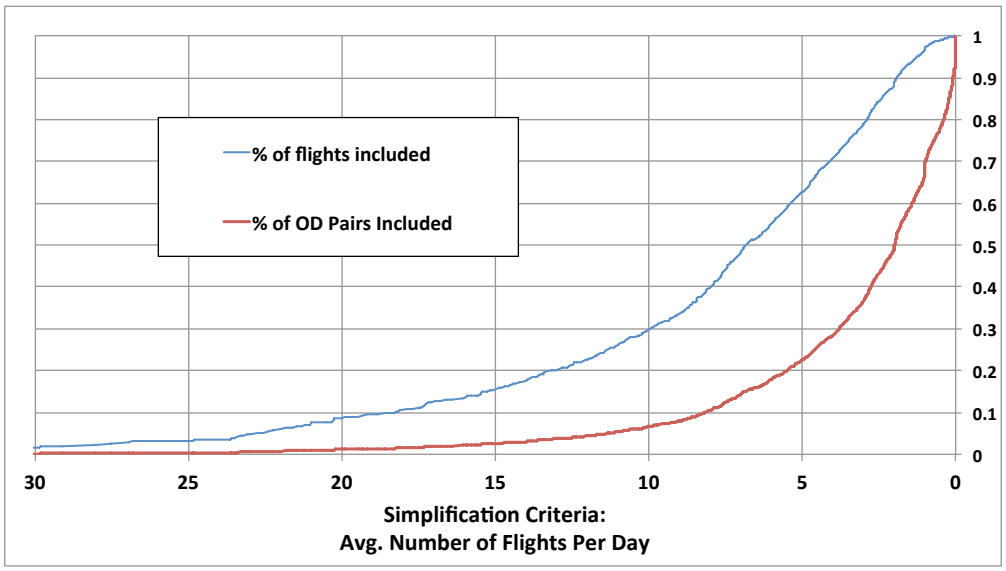


Figure 2-1: Percentages of flights and OD pairs included for different network simplifications.

Table 2.1 displays statistics associated with the top 20 most connected airports (according to their connectivity to other highly-connected airports). We anticipate the behavior of airports with a large number OD pairs in the simplified network will have the greatest influence on overall NAS delays due to their connectivity. We see that large airline hubs are at the top of the list. The last column in the table lists the number of connections each airport shares with the other airports in this table.

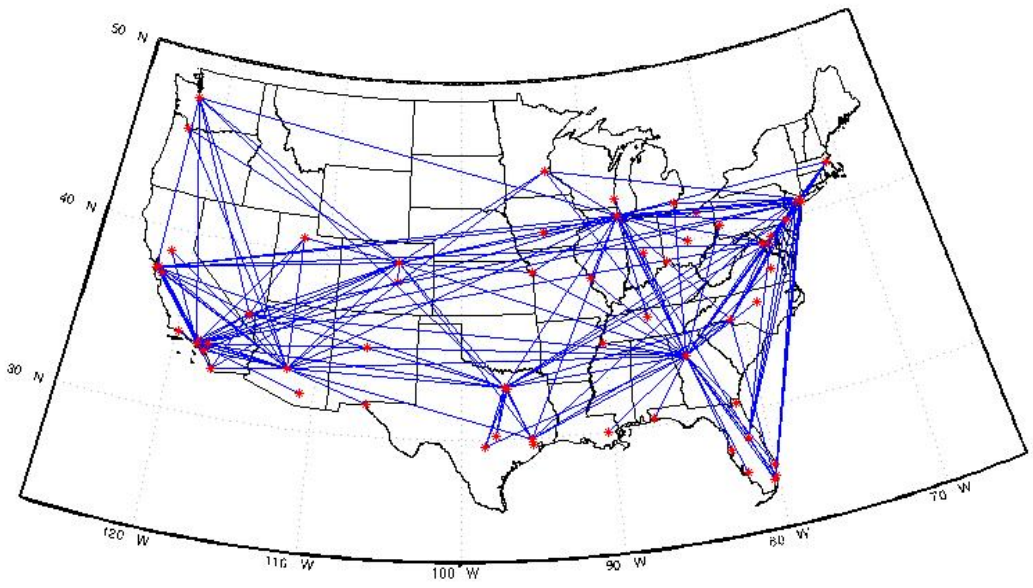


Figure 2-2: Simplified NAS network showing OD pairs with at least 5 flights a day.

2.2 Flight Data Aggregation

We aggregate flight data in the same manner as [10]. Instead of predicting individual flight delays, we aim to predict expected delay levels of different airpots and OD pairs in the network. Thus, we treat the delay state of an OD pair or airport to be an estimate of the delay that a hypothetical flight using that resource at that time will experience. For example, if the DEN-BOS departure delay state is 20 min at 5:00pm, it means that the estimated departure delay for any DEN-BOS flight taking off at 5:00pm is 20 min.

This flight data aggregation requires several steps. First, we remove any flights from the dataset that were cancelled or diverted. Then, we use a moving median filter to obtain an estimate of the delay for a specific OD pair. The delay state of an OD pair at time t refers to the median delay of all the flights that fall within a 2-hour time

Airport	# Departure Links	# Arrival Links	Connectivity among top 20
'ATL'	88	88	38
'DEN'	45	46	36
'DFW'	53	53	34
'ORD'	67	67	34
'PHX'	39	39	34
'DTW'	19	19	32
'LAX'	38	38	32
'BOS'	19	19	30
'EWR'	22	22	30
'IAH'	36	36	30
'MCO'	22	23	30
'CLT'	26	26	28
'LAS'	26	27	28
'MSP'	17	17	28
'PHL'	19	19	28
'SFO'	29	29	26
'IAD'	12	12	24
'MIA'	15	15	24
'DCA'	14	14	22
'SEA'	21	21	22

Table 2.1: Airports with the most links in the simplified network.

window beginning at time t. This low pass filter mitigates high frequency changes and outlying flights by calculating the median of the data points. We assume that there are meaningful changes in the delay states on an hourly basis, so we use a 1-hour step size for the moving filter, which leads to 17,544 observations for the 2011-2012 dataset. Each observation contains the expected departure and arrival delays on the 1,107 OD pairs in the simplified network.

This methodology gives an estimate for the delay state on an OD pair at time t; in other words, the expected delay of a hypothetical flight on the route at time t. To render the estimate, the filter relies on data points existing within the two-hour time window at time t. However, there are numerous hours in the dataset during which no flights took place for a specific OD pair. While we could assume the delay during such hours is zero, this is not always true: consider the case when the delay states immediately before and after the time period are very high. This would indicate that, if a flight had taken place during the time period, it also would have likely been highly

delayed. To resolve this problem, we linearly interpolate any periods with no flights by considering the outputs of the moving median filter for the nearest hours (before and after) during which flights did take place. We do not conduct interpolation if the period without flights lasts longer than 6 hours. We consider anything greater than 6 hours to be a sufficiently large gap in time such that interpolation is not appropriate. We also do not interpolate if the end of the day (4:00am) is included in the period without flights. We define the start and end of each day as 4:00am EST. We do so because this represents the approximate time during the night in which there is the least air traffic; consequently, delays are generally lowest during this time.

We find that approximately 60% of the time periods in the dataset contain at least one flight and thus do not require interpolation. 10% of the time periods do not contain any flights and were consequently interpolated. The remaining 30% of the time periods also do not contain any flights, but are not interpolated because the period of no flights either lasts longer than 6 hours or spans across 4am.

2.3 Airport Variables

We create airport delay variables for use in our prediction models in a manner similar to the OD pair delay variables. We define an airport departure delay at time t as the mean of the expected delay of every OD pair departing from that airport at time t. Similarly, the airport arrival delay at time t is the mean of the expected delay of every OD pair arriving at that airport at time t. Since we have 158 airports in the simplified network, this methodology creates 316 total airport delay variables since each airport has a unique variable associated with its estimated departure and arrival delays.

2.4 Cancellations

The BTS Transtats database includes flights that were cancelled. Approximately 2% of flights were cancelled in 2011-2012 [8]. We define cancellations variables for use

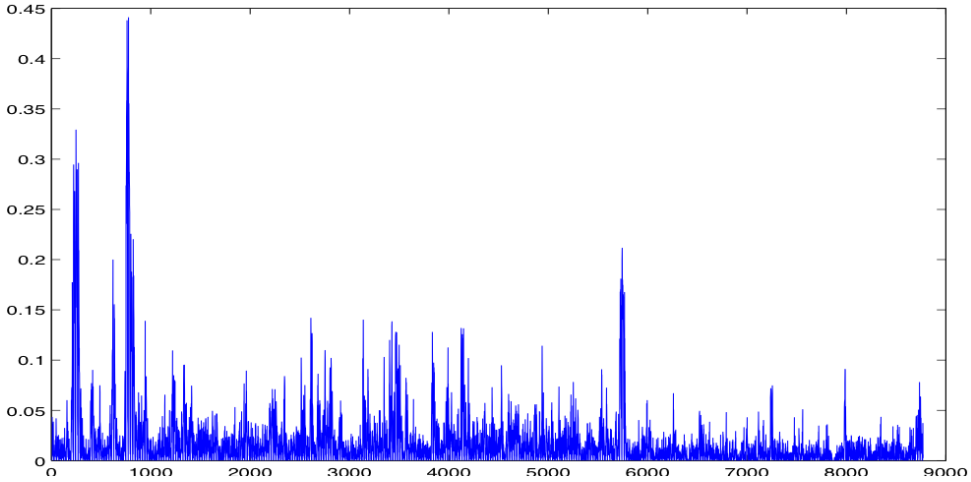
in our prediction models in a manner similar to the airport variables. Cancellations are binned by hour and departure airport using a two-hour window. No interpolation is performed on the cancellations data. For each hour, we divide the number of cancellations at a specific airport by the number of flights departing the airport in that hour. This yields an estimate of the percent of flights cancelled at an airport in a given hour. We use these estimates as cancellations variables; thus, we have 158 cancellations variables, each with 17,544 data points for the 2011-2012 dataset.

We assess these 17,544 cancellations points for seasonality since it is plausible that cancellations rates might be correlated with bad weather months. Figure 2-3 presents a time series of the percent of flights cancelled in each hour of our 2-year dataset. The time series appears random. Furthermore, the correlation between cancellations in 2011 and 2012 is -0.029, allowing us to conclude that seasonality does not significantly impact cancellations.

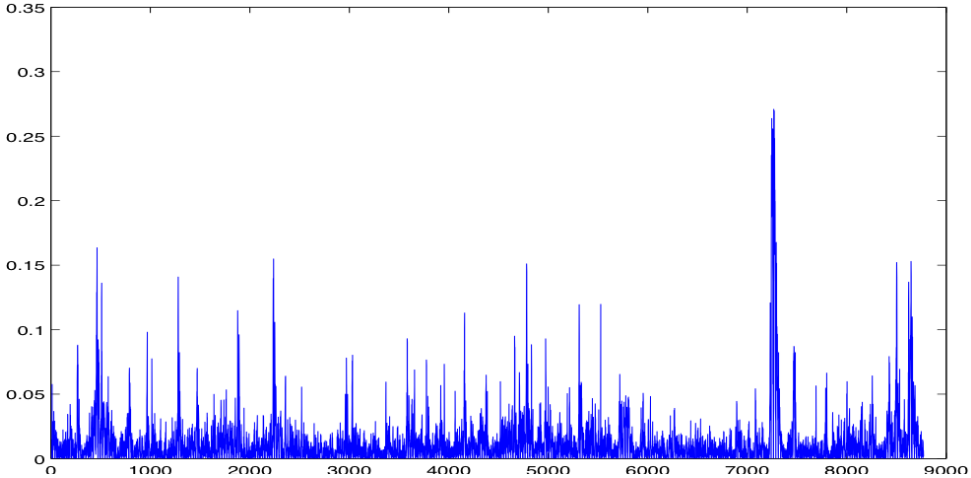
2.5 Analysis of Temporal Variables

The goal of this section is to identify time-related variables that can play an important role in predicting the level of delay of a certain link in the network. The analysis of the different variables presented in this section will focus on one specific link. We use ORD-EWR departure delays to evaluate the delay prediction model since ORD-EWR commonly experiences heavy delays and since it links two large airline hubs in New York City and Chicago.

Since we use a 1-hour time step in the dataset, each delay and cancellation observation can be linked to a specific hour in the day. Likewise, each observation corresponds to a particular day of the week and month of the year. We call these three categorical variables (time-of-day, day-of-week and month) 'temporal' variables, since they all relate to the timing of an observation. We hypothesize that one or more of these temporal variables may explain seasonality in the delay data. It is reasonable to assume that airline delays are greatest during periods of high volume, such as in the afternoons or on Mondays or Fridays.



(a) 2011



(b) 2012

Figure 2-3: Time series of cancellations percentages, binned by hour.

To test this hypothesis, we conduct ANOVA and multiple comparisons tests to evaluate the dependence of the departure delay with each of the three temporal variables. The methodology in this section is the same used in [10].

Because of the skew in the OD pair delay distributions, we use the Kruskal-Wallis ANOVA test for comparison of medians of non-normal data. We find that for every categorical variable, the p-value of the ANOVA test is less than 0.01. Thus, for each variable, we reject the null hypothesis and accept the alternative that at least one of the median delay values is different from another for different levels of the variable.

Extending this analysis further, we conduct multiple comparisons tests on every

variable to assess how each level of the variable differs from the other levels of the variable. We use the Tukey-Kramer significance criterion. Both time-of-day and day-of-week exhibit different delay values for many levels. Figures 2-4 and 2-5 display the estimates and comparison intervals for the time-of-day and day-of-week variables. 5:00am EST exhibits the lowest delay value; consequenlty, in subsequent chapters choose 4:00am EST to be the cutoff between operating days (5:00am represents the first hour in an operating day).

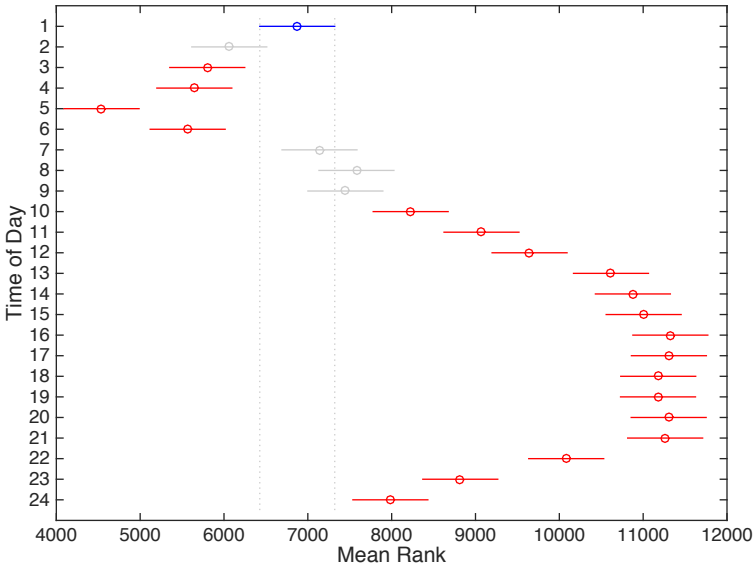


Figure 2-4: Time-of-day multiple comparisons test for ORD-EWR departure delays.

However, the month variable demonstrates some levels that are not significant when a pairwise comparison is conducted. Similar to [10], we choose to group the months into three different seasons: low (October-November), medium (January-May), and high (June-August, December). High volume and convective weather are the primary cause for higher delays during summer months, and high volume is also the cause for higher delays during December. Figure 2-6 displays the estimates and comparison intervals for seasons. Each season's delay levels are found to be statistically different in pairwise comparisons with the other seasons.

For the remainder of this thesis, we refer to the time-of-day, day-of-week, and season variables as the three temporal variables.

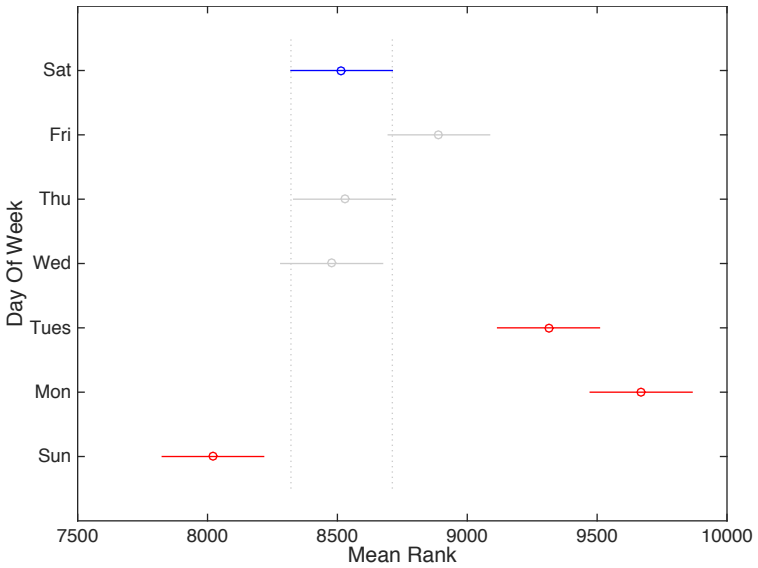


Figure 2-5: Day-of-week multiple comparisons test for ORD-EWR departure delays.

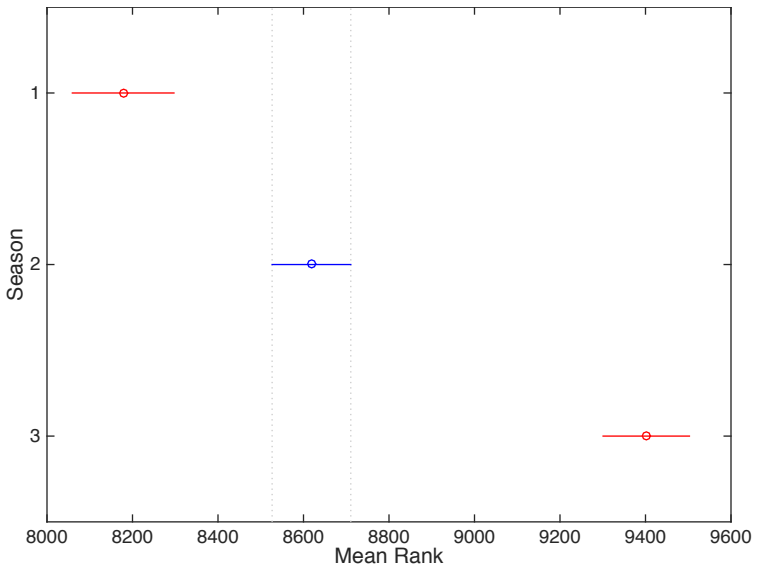


Figure 2-6: Tukey-Kramer multiple comparisons test for seasons. Note that we group months according to delay levels, not according to the gregorian calendar.

Chapter 3

Analysis of NAS Delays

The goal of this chapter is to classify the NAS network into groups of typical 'snapshots' that can describe the state of the U.S. airspace in a simple, concise manner using a single variable. To achieve this, we use k-means clustering to classify the state of the NAS in a given hour into a predefined number of groups. We define the NAS delay state at time t as a vector of the departure delays of all of the OD pairs at time t. The clustering algorithm outputs which 'typical state' is closest to each of the 17,544 observations, where the 'typical states' are given by the centroids of each of the clusters.

3.1 K-means Clustering Algorithm

The k-means algorithm is a centroid-based clustering algorithm. K-means partitions n observations into k clusters which minimize the within-cluster sum of squares. K-means minimizes the following objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^j - c_j\|^2$$

where x_i are the observations and the c_i centroids of each of the clusters.

3.1.1 Number of Clusters

To use k-means clustering, the number of clusters must be specified a priori. The performance of the k-means clustering algorithm when a specific number of clusters is used can be assessed by calculating the sum of the intra-cluster distances. As the number of specific clusters increases, the sum of the intra-cluster distances typically decreases, indicating a better fit to the data, but this often comes at the expense of simplicity. At the extreme, each observation becomes its own cluster. Figure 3-1 displays the sum of the intra-cluster distances for different numbers of clusters. The total intra-cluster distance decreases as the number of clusters increase. We choose to use 7 clusters for the remainder of the thesis for two reasons: first, the total intra-cluster distance for 7 clusters is relatively low and does not significantly decrease with additional clusters; secondly, 7 clusters allows for a reasonable representation of national delay centers since it includes clusters with high delays at or near New York City, Chicago, Atlanta, Texas, and San Francisco.

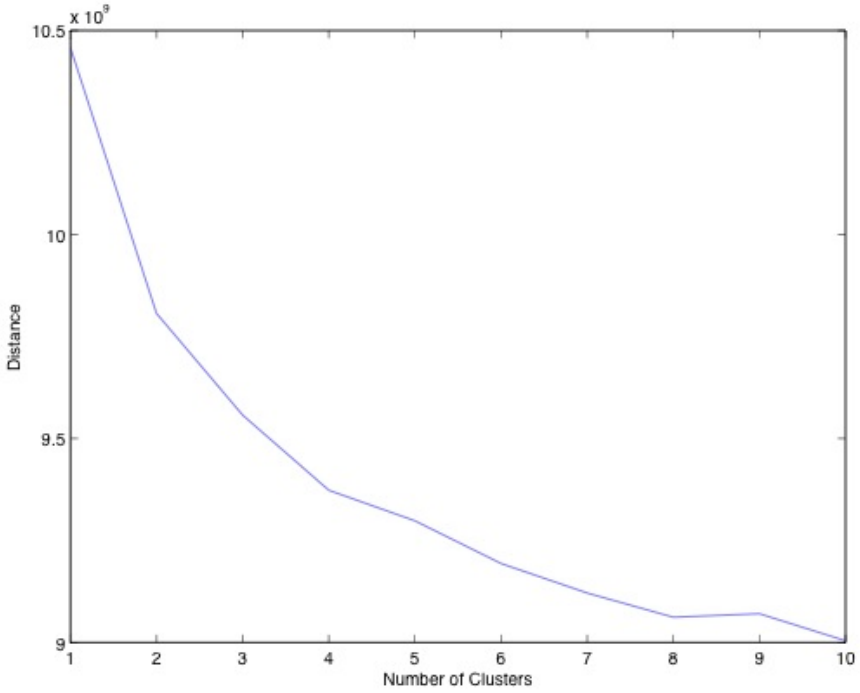


Figure 3-1: Total intra-cluster distance vs. number of clusters.

3.2 NAS State Cluster Centroids

Figure 3-2 displays each of the clusters' centroid delay values, with the color of the OD pair indicating the expected delay on that route. Warmer colors indicate a larger delay. It is readily apparent that the NAS states cluster delays according to regions, with the New York City, Chicago, Atlanta, Texas, and San Francisco areas all being highlighted. Thus, we can observe that delays are highly interdependent on regional behavior, as we would expect.

Table 3.1 provides additional metrics regarding each NAS state. Clusters 5 and 7, the 'Low NAS' and 'Medium NAS' states, occur much more frequently than the other clusters. The 'Low NAS' state also occurs, on average, for almost 13 hours. This is reasonable, since we would expect the default state of the NAS to be one with very few delays. This state occurs at least once in every day within our two-year dataset, indicating that in the middle of the night the NAS always "resets" back to a low delay state. In contrast, the states representing high delays in specific areas, such as the 'High ATL' state, occur infrequently. Only 5% of days experienced a 'High ATL' hour, and on average this state only lasted 5 hours, indicating the delays in Atlanta tend to resolve themselves slightly faster than delays at other hubs such as ORD. The average OD pair delay during each state reveals the severity of the NAS state. OD pairs during the 'High ATL' state exhibit average expected delays over seven times greater than during the 'Low NAS' state in cluster 7. We also observe that the 'High SFO' state is only slightly worse on a nation-wide scale than the 'Med NAS' state.

Cluster #	Number of Hours (17544 total)	Percentage of Days Each State Occurs	Avg. Active Time (Hours)	Average OD Pair Delay (min)	Specific Delays
1	1567	33%	6.29	10.6	High SFO
2	714	13%	6.93	18.4	High ORD
3	289	6%	6.15	18.1	Med-High TX
4	191	5%	5.16	24.0	High ATL
5	5021	89%	5.55	9.5	None
6	719	15%	6.20	20.8	High NYC
7	9043	100%	12.90	2.6	None

Table 3.1: NAS delay states statistics.

Figure 3-3 plots the frequency with which each state occurs for each hour of the day. Around 6:00am EST, every day is in a low NAS state. Less than 10% of days remain at a low NAS state throughout the day. Instead, over half of the days transition into a medium NAS state by the afternoon, and the rest experience one or more of the high delay states in the latter half of the day. This exemplifies a typical trend in airline delay propagation in that delays that originate during the middle portion of the day tend to propagate and manifest themselves into the evening hours.

Figure 3-4 is a histogram of the number of hours each state occurs, separated by month. In general, the summer months experience a greater number of highly delayed states, particularly for NYC, ATL, and ORD. The high SFO state is the one exception to this; it occurs most frequently during the winter months. Its decreased prevalence during the summer months is most likely a due to an increase in the delays at the NYC, ATL, ORD, and TX airports relative to SFO rather than an absolute decrease in delays at SFO.

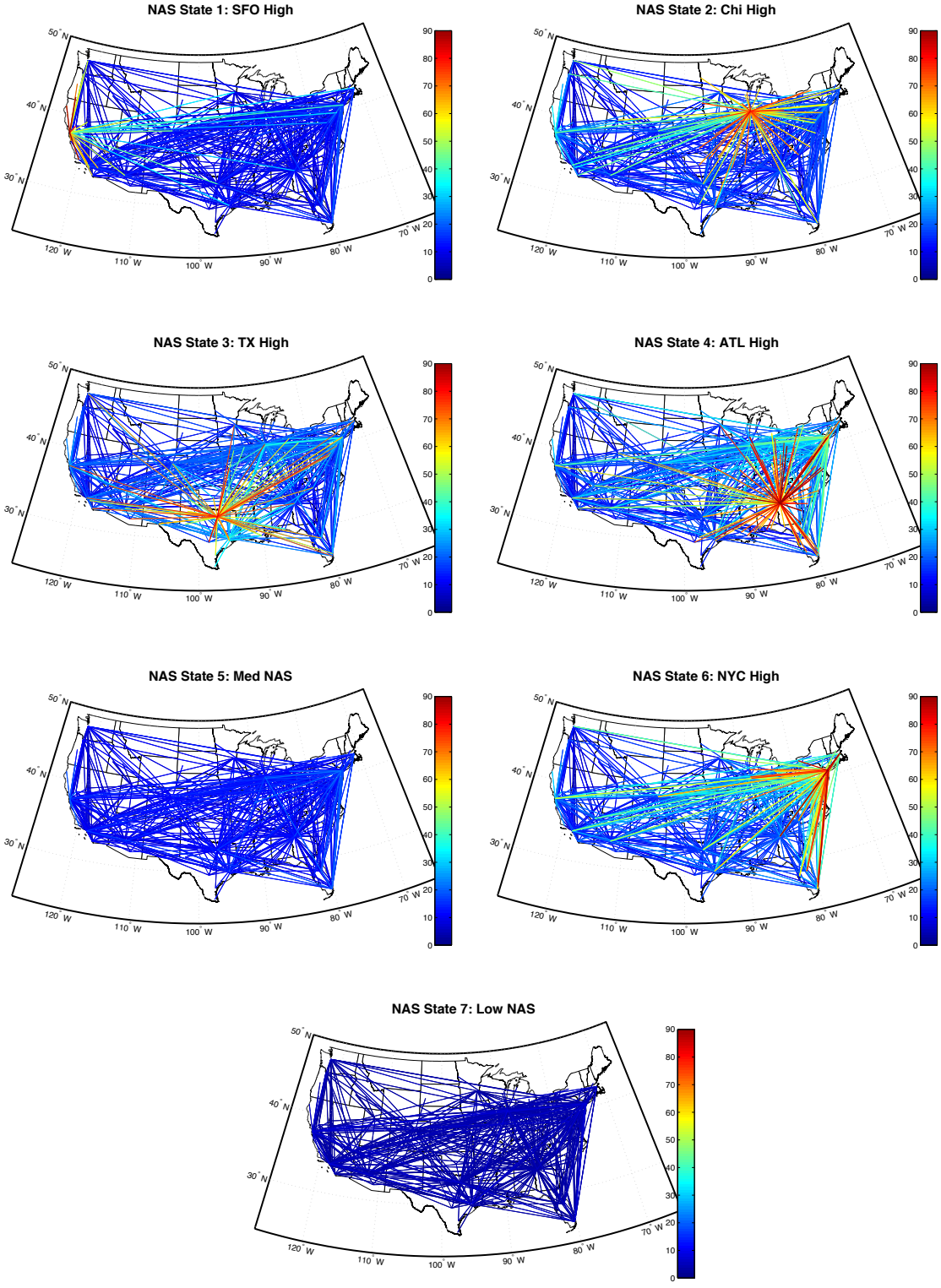


Figure 3-2: Centroids of NAS delay states for seven clusters.

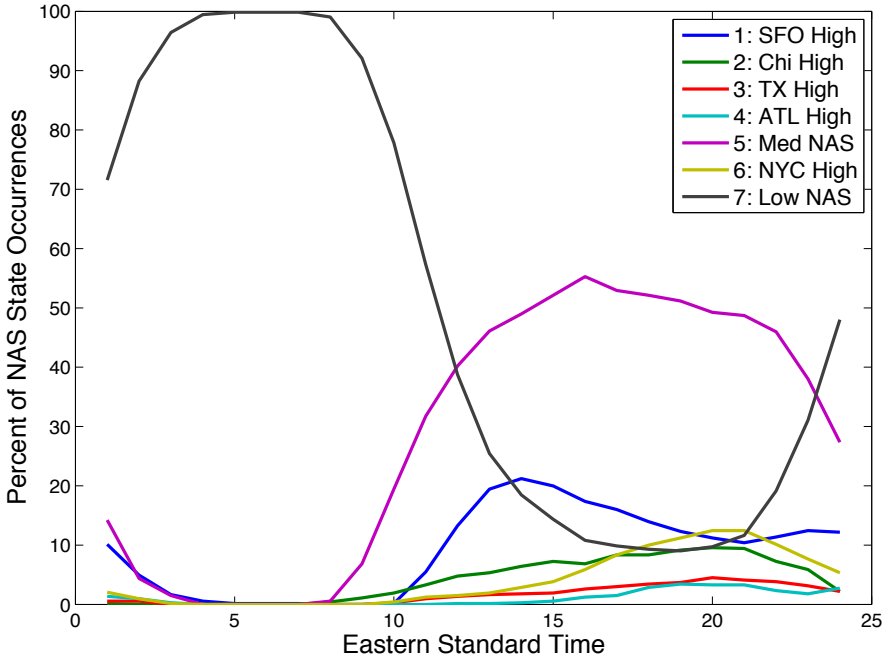


Figure 3-3: Frequency of NAS state occurrences.

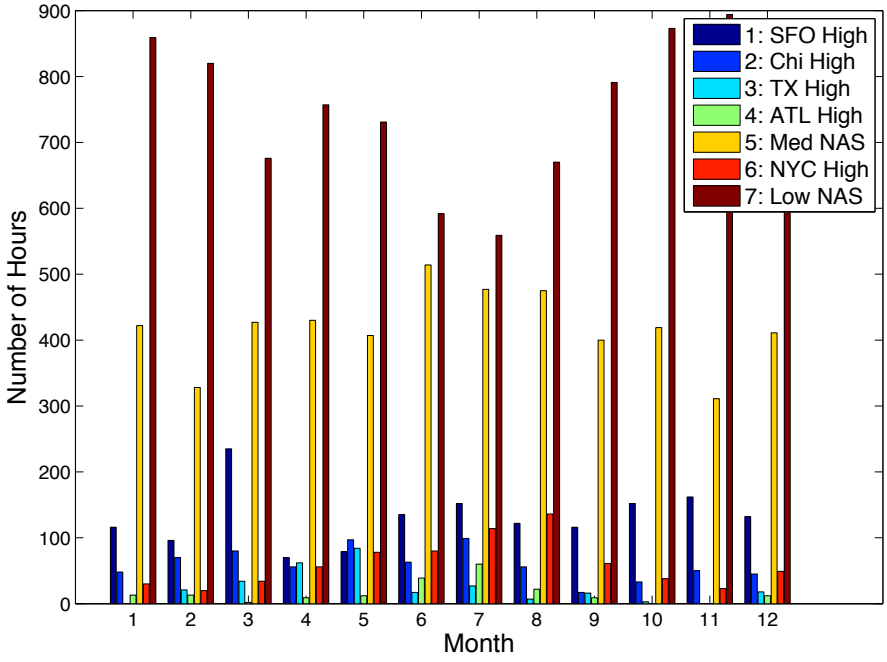


Figure 3-4: Hourly occurrences of NAS state; separated by month.

Chapter 4

NAS Type-of-Day

4.1 NAS Type-of-Day Clustering

We also define a NAS type-of-day variable with the goal of describing an entire day's worth of U.S. airline activity in a single metric. To do so, we cluster entire days of OD pair delay data in much the same way we do for the NAS delay states in Chapter 3. For the NAS type-of-day, however, we have 731 observations, one corresponding to each day in the 2011-2012 dataset. Each of these 731 data points contains 1107x24=26568 variables (Number of OD pairs x 24 hours, we have one observation per hour due to the 1-hour step size of the moving median filter). We define the start and end of each day as 4:00am EST. We do so because this represents the approximate time during the night in which there is the least air traffic; consequently, delays are generally lowest during this time.

Figure 4-1 shows the total intra-cluster distance for different numbers of clusters using the k-means clustering algorithm. We followed the same methodology presented in Chapter 3 to choose the number of NAS state clusters (based on distance reduction and qualitative description of the centroids), and we chose seven clusters again.

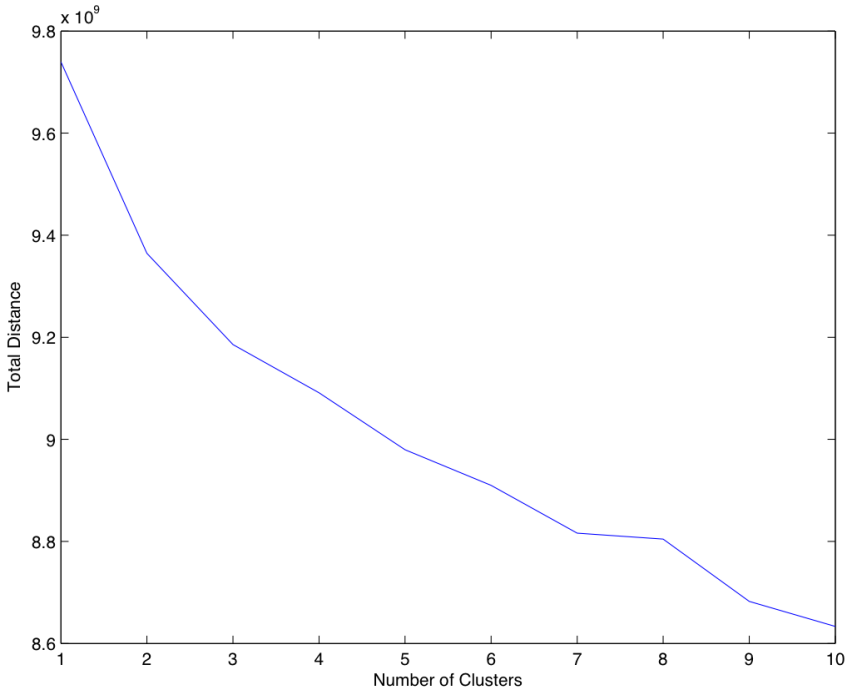


Figure 4-1: Total intra-cluster distance vs. number of clusters.

We also perform the same clustering using both OD pair departure and arrival delays (each observation has 1107x2x24=53136 variables). However, since OD pair departure and arrival delays are highly correlated, adding arrival delays in the clustering algorithm has a trivial affect on the resulting cluster centroids, so we retain our clusters found by just using OD pair departure delays.

Figure 4-2 displays each of the type-of-day clusters' centroid delay values during each cluster's most heavily delayed hour in the day. From these mappings we deduce that the type-of-day clusters are very similar to the NAS state clusters, in that both group largely according to regional delays. The same delay epicenters (NYC, ORD, ATL, TX, SFO) that were present in the NAS state clusters are also present in the type-of-day clusters. Table 4.1 provides additional metrics regarding each type-of-day cluster. We note that over 50% of days are considered to be a 'Low NAS' type-of-day.

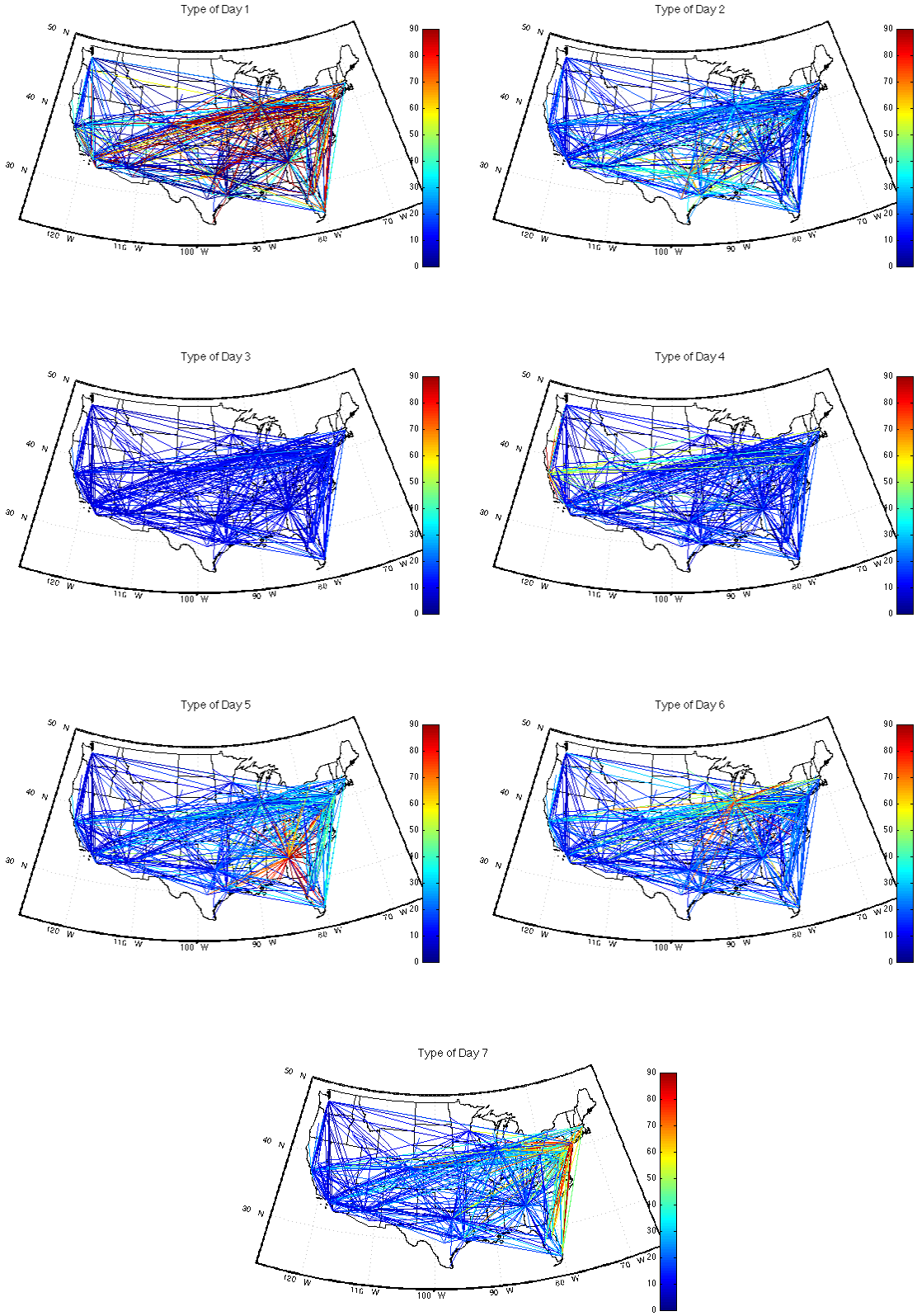


Figure 4-2: Centroids of the seven type-of-day clusters during each cluster's most heavily delayed hour in the day.

Cluster #	Number of Days (731 Total)	Average Centroid Link Delay (min)	Qualitative Description
1	2	42.7	High NAS
2	29	22.0	High TX
3	458	9.9	Low NAS
4	82	14.7	High SFO
5	27	26.1	High ATL
6	60	21.5	High ORD
7	73	23.5	High NYC

Table 4.1: NAS type-of-day clustering.

Figure 4-3 displays the number of occurrences of each of the type-of-day clusters by month during the 2011-2012 timeframe. Similar to Rebollo de La Bandera, we find that the highest delay days occur most frequently during the summer months. In particular, the type-of-day corresponding to high delays around New York City airports occurs almost twice as frequently during the summer months than during the other months of the year. Besides the low NAS type-of-day, in which no significant delays exist, the most frequent type-of-day corresponds to high delays at SFO. This represents a drastic change from the 2007-2008 results from Rebollo de la Bandera in which SFO didn't even appear as a unique type-of-day. We have several explanations for this behavior. First, since 2008, air traffic volumes have increased more in the western half of the United States than in the East. An increase in delays has likely paralleled an increase in volume for airports such as San Francisco. Second, it is noteworthy that the months that see the greatest number of occurrences of the high SFO type-of-day are months that have relatively low numbers of occurrences of high ATL, high ORD, and high NYC delays. Thus, another explanation for the frequent appearance of the high SFO type-of-day is that it often replaces the low NAS type-of-day under conditions of moderate delays, particularly near the west coast. However, whenever the NAS becomes more heavily delayed, the high ATL, high ORD, and high NYC clusters become much more representative of the NAS; consequently, these high delay days quickly replace the high SFO type-of-day. The high SFO type-of-day is essentially a 'weak' cluster; it likely accounts for many uneventful days that simply have a higher than normal NAS delay. This is reflected in the average centroid link delay for the different type-of-day clusters: for the high SFO cluster, this metric is

only 14.7 min; a mere 4.8 min more than the low NAS cluster. All other clusters have an average link delay greater than 20 minutes.

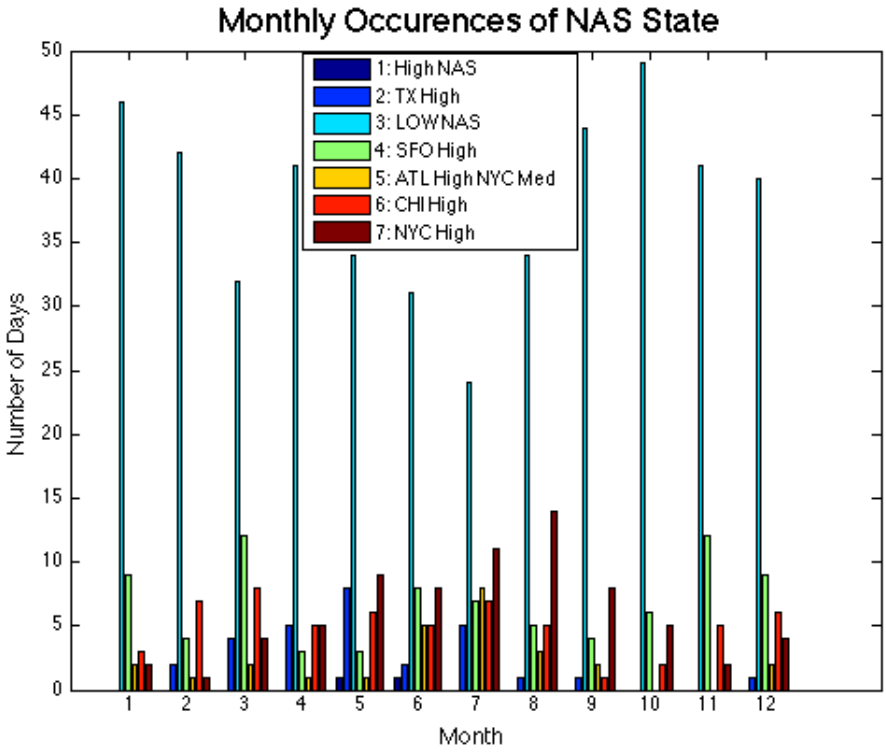


Figure 4-3: Monthly occurrences of NAS type-of-day.

In order to further investigate the differences between type-of-day clusters, we aggregate the expected delays on each OD Pair during each hour of the day for each type-of-day (each type-of-day has 1107 OD pairs and 24 hours, resulting in 26568 OD pair-hours). We then plot a histogram of the 26568 OD pair-hours for each type-of-day in Figure 4-4. We observe the general exponential distribution of delays that was previously described: hours of low-delay are much more common than hours of high delay. As expected, virtually no OD pairs are ever delayed more than 30 minutes during the low NAS type-of-day. Conversely, the high NAS type-of-day displays a much longer tail: nearly one in four flights is delayed greater than 30 minutes. It is interesting to note, however, that the high NAS type-of-day also has a greater amount of low-delay OD pair hours than the other high delay types of days. In comparing the low NAS type-of-day to the high SFO type-of-day, we find that the high SFO

type-of-day has a greater amount of flights delayed greater than 10 minutes and a much longer tail than the low NAS type-of-day. The other four high delay states (TX, ATL, CHI, NYC) all exhibit very similar delay distributions.

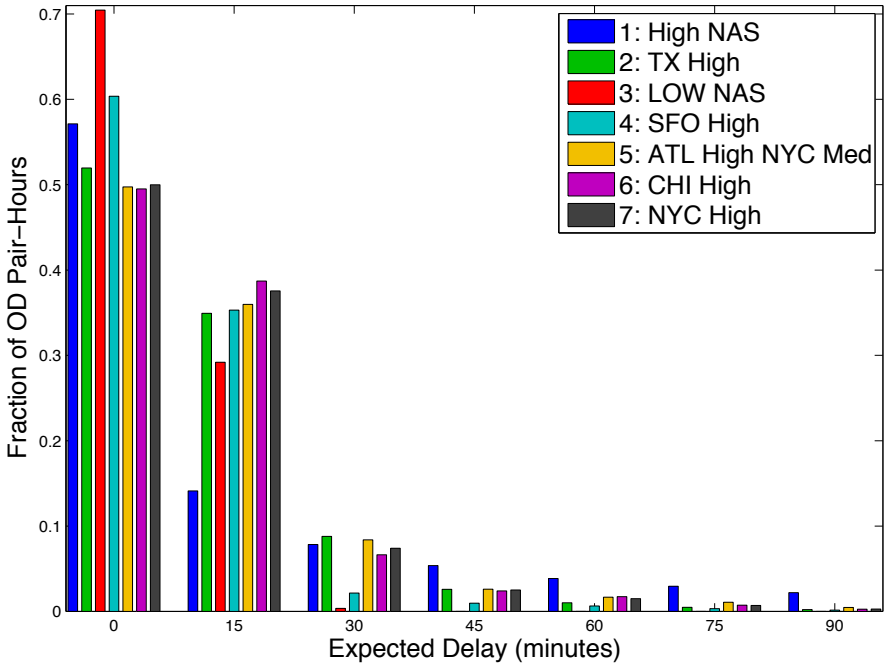


Figure 4-4: Histogram of the expected delay on OD pairs for different types of days.

Figure 4-5 displays the average delay of major U.S. airline carriers for each type-of-day. As expected, the low NAS type-of-day exhibits the lowest delays for all carriers. Conversely, the high NAS type-of-day exhibits the highest delays for several carriers. The chart reveals that many of the types of days are attributable to a specific airline that operates heavily in the afflicted area. For example, Delta experiences the highest average delay during the high ATL type-of-day, which is plausible since ATL is Delta's primary hub. Similarly, JetBlue's highest average delay occurs during the high NYC type-of-day.

The type-of-day clustering described thus far used only the expected departure delay of OD pairs. We performed the same methods using both expected departure delays and expected arrival delays. Thus, for each hour, there are two variables associated each of the 1107 OD pairs in the simplified network, resulting in 2214 values. We re-run the k-means clustering on the 731 days in the two-year dataset, with

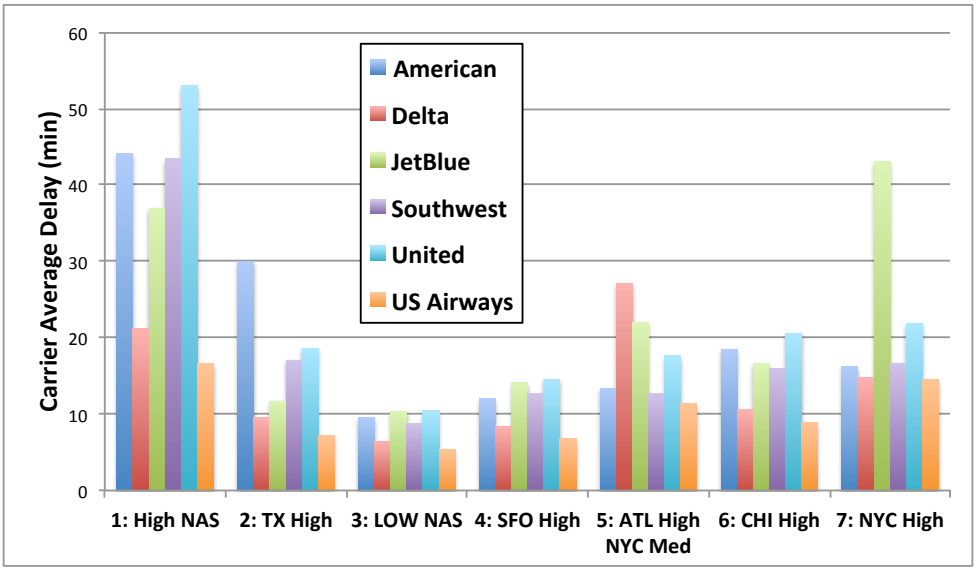


Figure 4-5: Average carrier delay for each NAS type-of-day.

each day having 2214*24 = 53136 values. The resulting clusters are nearly identical to the clusters obtained using only departure delay values. This is to be expected for two reasons: 1) the departure delay values are included in both methods; and 2) the expected arrival delay values for OD pairs closely mirror their respective departure delay values. Thus, included the arrival delay variables introduces a trivial amount of new data to the algorithm. Because the two clustering methods yield nearly identical clusters, for simplicity we choose to only use departure delays in the NAS type-of-day clustering.

4.2 NAS Type-of-Day Prediction

4.2.1 Predicting NAS Type-of-Day Using the Elapsed OD Pair Delay Data

With our given definition of the NAS type-of-day, we cannot conclusively say what the current NAS type-of-day is until the day terminates. In this section, we develop the capability to predict the current NAS type-of-day using only the subset of hours in that day that have already elapsed. Later in this thesis we include the predictions

for the NAS type-of-day in the OD pair delay prediction models and investigate the effect of this addition.

The NAS type-of-day was created using k-means clustering on the expected delay of each OD pair in each hour of the day. We use a random forest model to predict the NAS type-of-day in each hour of the day using only the hours of the day that have already passed. We do so by identifying which cluster centroid is closest (using Euclidian distance) to the subset of hours that have already passed. For example, to predict the NAS type-of-day at 10:00am, we find which type-of-day cluster centroid has delays from 4:00am to 10:00am that are closest to the 6 hours in the day that have elapsed since 4:00am. Thus, for each day in our two-year dataset, we have 24 separate predictions for the NAS type-of-day. The accuracy of this prediction method, labelled "Nearest cluster", is plotted in Figure 4-6.

We then run a series of random forest models to improve upon the nearest cluster method for predicting the NAS type-of-day. The first random forest model simply uses the nearest cluster predictions for each of the elapsed hours in the day as the predictor variables. Thus, when predicting the type-of-day at 10:00am EST, 6 hours have elapsed since the start of the day at 4:00am, so there are 6 predictor variables included in the model, with each corresponding to the nearest cluster prediction after a certain hour has elapsed. Since are 24 hours in the day, we run 24 separate random forest models, one for each hour in the day. The results of this method, labeled "RF-Nearest cluster" are also plotted in Figure 4-6. This random forest model significantly improves the accuracy of the type-of-day prediction in the middle part of the day when delays first start to occur.

The third NAS type-of-day prediction model we create improves upon the second by including the NAS type-of-day of the previous day as a predictor variable in the random forest model. As shown in Figure 4-6, this addition improves prediction accuracy in the earlier hours of the day at which time there is less data from the current day to make nearest cluster predictions.

We also see that as the day progresses, the previous type-of-day becomes less important, and eventually the predictive capability of the nearest clusters significantly

outweighs that of the previous type-of-day, thus causing the accuracy of the second and third models to converge by approximately 6:00pm EST.

The fourth and final model we create expands upon the third by also including the day of week and season temporal variables in the random forest model. Again, Figure 4-6 shows that the added variables improve the NAS type-of-day prediction accuracy, particularly near the beginning of the day.

For the remainder of the thesis, we choose to use the fourth model for predicting the NAS type-of-day. We use 30 trees and all 17,544 unique hours in the dataset to train the random forest model.

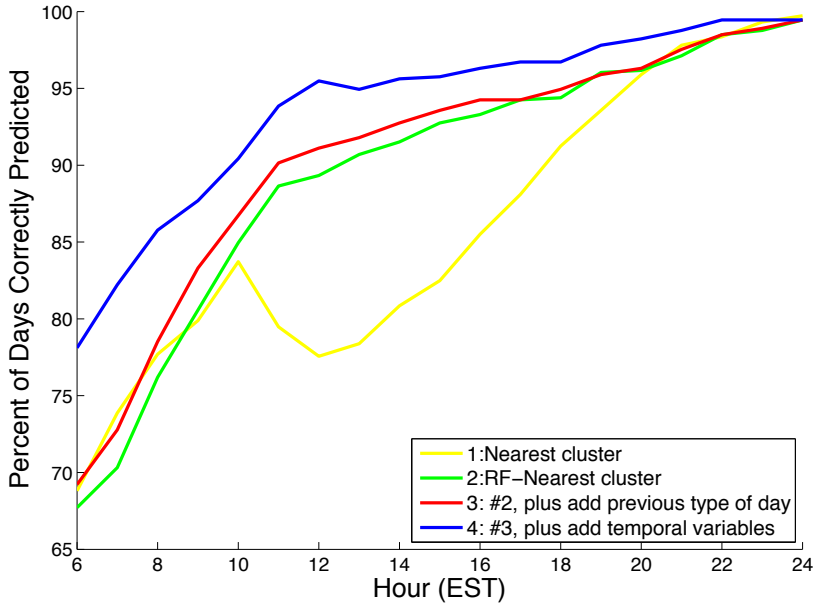


Figure 4-6: NAS type-of-day prediction accuracy of different prediction models.

Overall, using this fourth model yields an prediction accuracy of 93.7%. Figure 4-7 displays the performance of the model on each type-of-day individually. We notice that the model has a very high prediction accuracy when the actual type-of-day is 'Low-NAS'. This occurs because the 'Low-NAS' type-of-day is the most common, and thereby the default, type-of-day (approximately half of the days in the dataset are 'Low-NAS'). Thus, whenever the model incorrectly predicts the type-of-day, it usually incorrectly predicts a 'Low-NAS' day when in reality the day eventually reflects one of the six high delay clusters. The 'High NAS' type-of-day exhibits a discrete behavior

in this graph since there are only two days in the entire dataset that are 'High NAS'. For both days, by 9:00am EST, the model correctly predicts the type-of-day. As shown in Figure 4-7, the prediction accuracies of the other types of days follow the same general increasing trend as the day progresses.

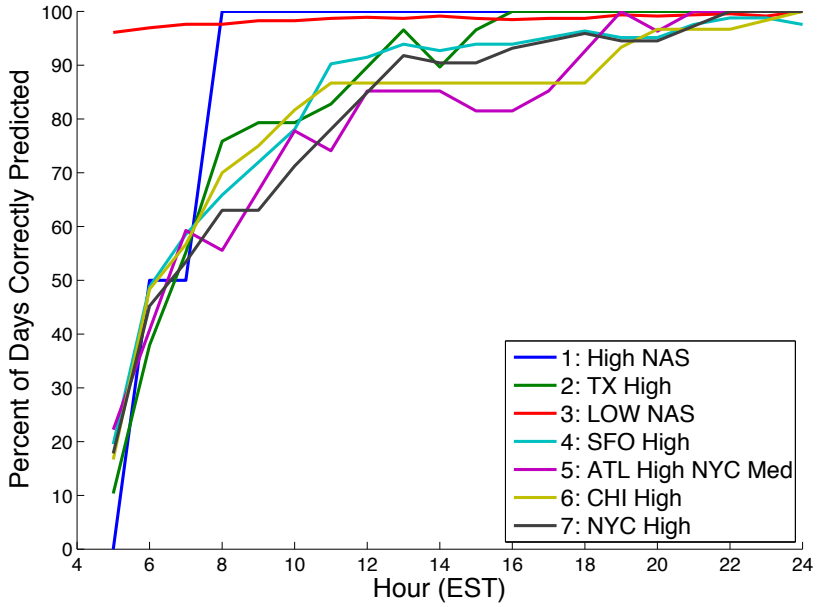
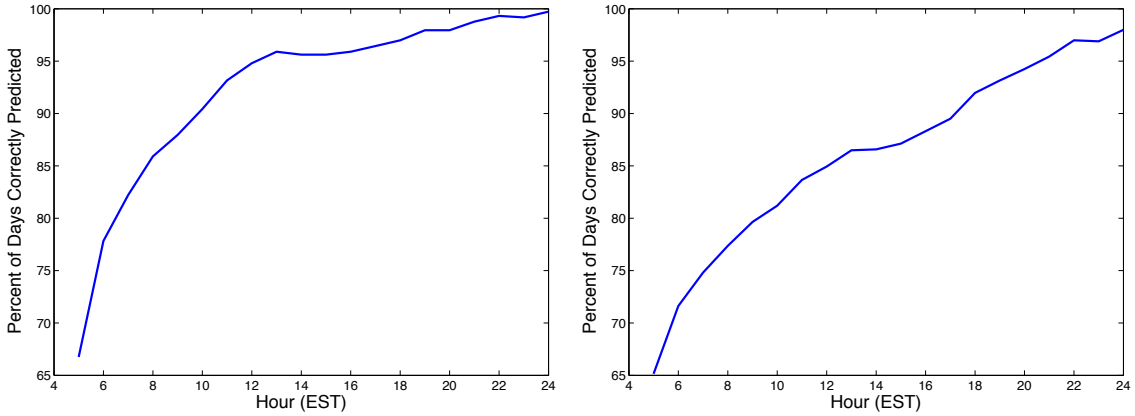


Figure 4-7: Accuracy of type-of-day predictions by hour, separated by actual type-of-day.

4.2.2 2013 Type-of-Day Prediction Performance

We also investigate how well our model for predicting the NAS type-of-day described in Section 4.2.1 performs on data from 2013. We continue to use the type-of-day prediction model that is trained on 2011-2012 data, but we assess its accuracy in predicting the NAS type-of-day for 2013. Figure 4-8 plots the accuracy of the type-of-day predictions for 2013 as the day progresses. As shown in Figure 4-8, the baseline accuracy (at 5:00am EST) for predicting the type-of-day in 2013 is approximately 65%, which is slightly lower than the baseline accuracy when predicting type-of-day in 2011-2012. By noon (EST), the accuracy for predicting type-of-day in 2013 is approximately 85%. While this accuracy is slightly lower than that obtained from predicting on 2011-2012 data, a small decrease in accuracy is reasonable when testing

on a new dataset. Nevertheless, a 2013 accuracy of 85% by noon EST indicates that the type-of-day prediction model performs well both on the original 2011-2012 data used for training and the new 2013 data.



(a) 2011-2012 (b) 2013

Figure 4-8: Accuracy of type-of-day predictions, by hour, for the 2011-2012 and 2013 datasets.

Chapter 5

OD Pair Delay Prediction Models

In this chapter we evaluate different classification and regression delay prediction models with the goal of determining which model is most appropriate. We use many of the same models as [10]. We use ORD-EWR departure delays to evaluate the delay prediction model since ORD-EWR commonly experiences heavy delays and since it links two large airline hubs. We choose a classification threshold of 60 minutes, which is a binary indicator of whether the expected delay on the link is greater than or less than 60 minutes. Additionally, we choose a prediction horizon of 2 hours which signifies how far into the future we make predictions. For example, with a 2 hour prediction horizon, if we are predicting ORD-EWR departure delay at 5:00pm, we use the 3:00pm OD pair delays, airport delays, and NAS state as explanatory variables. We use actual gate departure and arrival times for this analysis.

5.1 Training and Test Sets

For the remainder of this thesis, except where indicated otherwise, training sets of 3,000 points and testing sets of 1,000 points are used when fitting the delay prediction models. Analysis of the effect of the training set size is conducted in the next chapter. We use repeated random sub-sampling, also known as Monte Carlo cross-validation (MCCV) by fitting and testing each model we create on 5 unique training and test set pairs. Each training set is used to fit a model, which is then assessed for accuracy

using the corresponding test data. The results are averaged over the 5 splits. This allows for a measure of the error variability and a better estimate of the test error.

All of the training and test sets are created using over-sampling, which deliberately selects rare data points in order to obtain more precise estimates of rare events. Since the delay data is highly non-normal, hours of very large delay are relatively rare; therefore, if purely random sampling was used, these large delay events might never be chosen in training and testing sets. Instead, by using over-sampling, we achieve a balanced dataset that contains equal amounts of high and low delay data points. Histograms of the departure delays on the ORD-EWR link before and after oversampling are show in Figure 5-1 and Figure 5-2. The oversampling shown in Figure 5-2 allows for a greater percentage of high delay points in the training set.

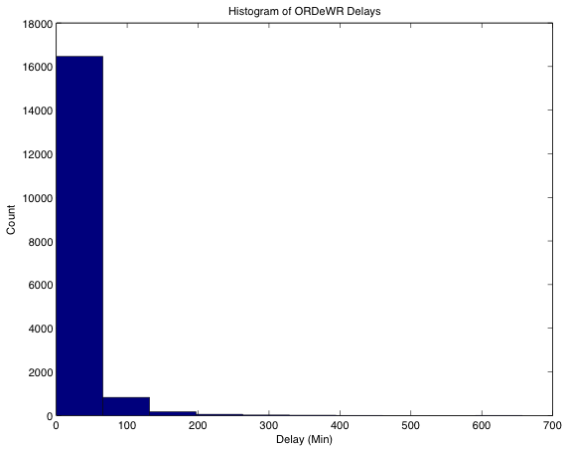


Figure 5-1: Histogram of ORD-EWR departure delays.

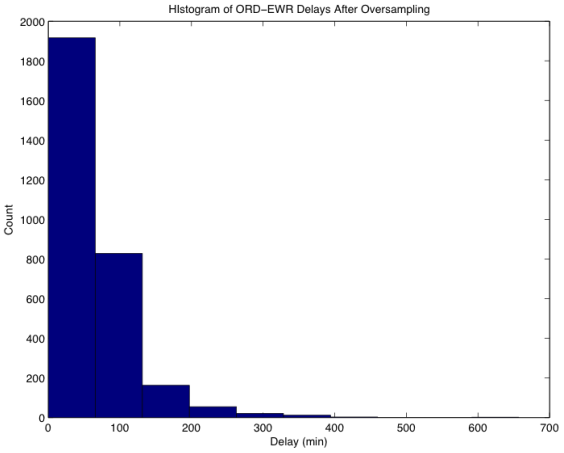


Figure 5-2: ORD-EWR departure delay training set after oversampling.

In addition to oversampling on these training and test sets, we choose to remove any interpolated data points with zero delay. We do so for two reasons: first, many of these points occur at night. Second, many of the zero delay points are hours in which there were no recorded flights within the two-hour time window, and the interpolation method that was previously described returned an expected delay of zero. However, there are situations in which, had there been a flight during that time, the delay on the flight would have been high; even with interpolation we simply don't have enough continuous flight data to express the high delay.

5.2 Explanatory Variable Selection

In the simplified network, we are have 158 airports, which results in 316 possible airport delay variables (158 for departures and 158 for arrivals). Similarly, we have 1,107 OD pair variables which results in 2,214 possible OD pair delay variables and 158 cancellations variables. Prior to creating the delay prediction models, we reduce the number of explanatory variables we use by selecting only the top 10 most influential OD pairs and the top 10 most influential airports (later in this thesis, when we include cancellations, we follow the same process).

We use the random forest regression algorithm to select variables for inclusion in the delay prediction models. We do so based on the variable's importance level in the random forest model. Our choice of random forest is twofold: 1) random forest has been demonstrated to be effective when the number of variables is large compared to the number of observations in the training set; and 2) because of the nature of the random forest algorithm, the importance of each variable is easily obtained. We use the following methodology to identify the set of OD pairs and airports of interest (the methodology is performed twice, once for OD pairs and once for airports:

1. Sampling 2,500 training data points from the data and fitting of a RF with 6 trees.
2. Selection of the 100 most important variables using the RF information obtained in the previous step.
3. Detailed analysis of the 100 most important variables: sampling of 3 different training data sets with 2,400 samples each, and fitting of a RF with 15 trees to each of those training data sets. The final variable importance values are the average of the values obtained from the 3 RFs.
4. The top ten OD pairs and airports are chosen according to the averaged variable importance values.

The results of this variable selection methodology for ORD-EWR departure delays are shown in Table 5.1.

Airport	Type of Delay	Normalized Importance
ORD	Departure	92.7
MDW	Departure	89.6
BNA	Departure	76.9
BWI	Departure	76.6
CLE	Departure	75.9
STL	Departure	73.7
PHL	Departure	72.5
IAH	Departure	71.4
DTW	Departure	68.2
ORD	Arrival	66.4

OD Pair	Type of Delay	Normalized Importance
ORDEWR	Departure	100.0
MDWEWR	Departure	42.0
STLEWR	Departure	38.7
BWIBNA	Departure	37.9
MDWBWI	Departure	36.5
BOSEWR	Departure	36.3
CLTEWR	Departure	34.6
MDWPHL	Departure	34.4
BWIRDU	Departure	34.4
ORDMIA	Departure	34.3

Table 5.1: Top 10 most influential airports and OD pairs for ORD-EWR departure delay prediction.

5.3 Classification Models

In this section we fit different classification models that predict ORD-EWR departure delays and evaluate their performance. For classification, we convert the expected ORD-EWR delay variable, obtained from the interpolation process described above, into a binary variable that indicates whether the expected ORD-EWR departure delay is high or low. In this section we use a classification threshold of 60 minutes to delineate between high and low delay levels; however, later in this thesis we consider different threshold values. We use a 2-hour prediction horizon, which is also varied later in the thesis.

The first classification model we consider is logistic regression. In order to use logistic regression, we first convert all of the categorical variables into binary variables. The mean error for the 5 fitted logistic regression models (MCCV) was 16.40% and the mean AUC value was 0.91. A brief inspection of the p-values resulting from the t-test for significance of explanatory variables indicates that the temporal variables, the NAS state, and several airport variables are likely significant in models of this type. The OD pair explanatory variables generally appear less significant, but still relevant. In particular, the OD pair variable corresponding to the route being modeled, in this case, the ORD-EWR variable (2 hours prior), is significant.

The second classification model we fit is a single classification tree. We grow models using the GINI index. We investigate pruning to prevent misclassification

due to either lack of fidelity or overfitting. Figure 5-3 shows the MCCV error rate for different pruning levels (the higher pruning level, the fewer nodes in the tree). We observe that the error rate is nearly monotonically increasing as we increase prune level. Because of this, we choose not to prune the classification tree. The mean error for the 5 fitted classification trees (MCCV) is 13.80%, and the mean AUC is 0.92. This represents a small increase in performance over the logistic regression model.

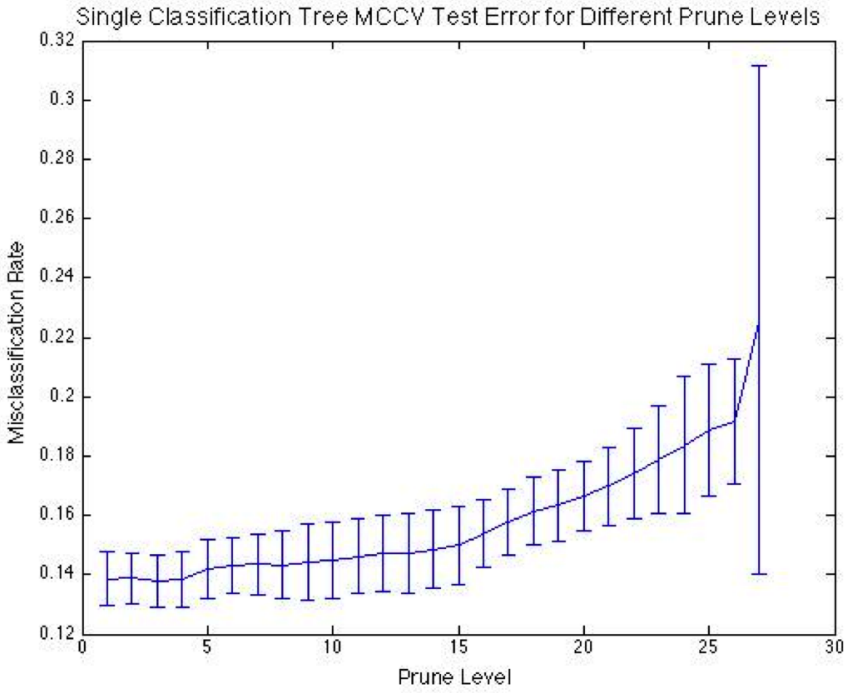


Figure 5-3: Single tree classification test error for different prune levels.

The third classification model we consider is random forest to build an ensemble of classification trees. For the random forest algorithm, we set the minimum number of observations per leaf to be 1 and the size of the random subset of variables searched at each split to be the square root of the total number of variables. Figure 5-4 shows the performance of the random forest model for different numbers of trees. While the misclassification error continues to decrease as the number of trees increases, it does not do so significantly after approximately 40 trees. Consequently, in the remainder of this thesis, when we use random forest for delay prediction models, we use 40 trees. The mean error resulting from the 5 random forest models is 8.40% and the AUC is

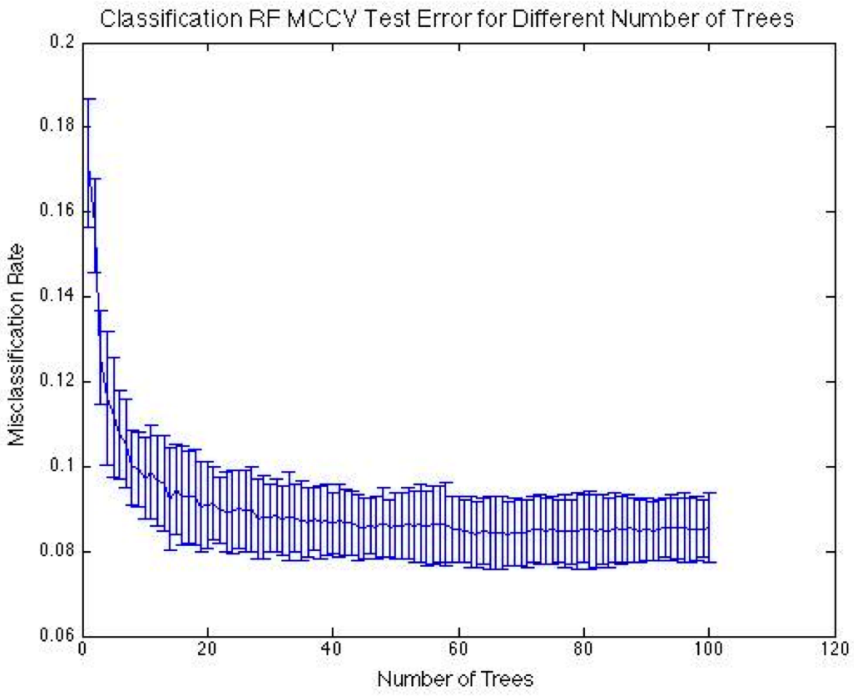


Figure 5-4: Random forest classification error for different numbers of trees.

0.97. These values indicate that, for classification, random forest is a better model than classification tree and logistic regression.

Figure 5-5 displays the ROC curves for the three classification models. As with the mean error values, random forest displays the best-performing ROC curve.

5.4 Regression Models

In this section we fit different regression models that predict ORD-EWR departure delays and evaluate their performance. The output of these models are continuous and represent the predicted delay on the link in minutes. Similar to the ORD-EWR classification models, we use a 60 minute classification threshold (which is necessary only to conduct oversampling when creating the training and test sets) and a 2-hour prediction horizon.

The first regression model we consider is linear regression. The significances of explanatory variables in this model are similar to those of logistic regression; that is,

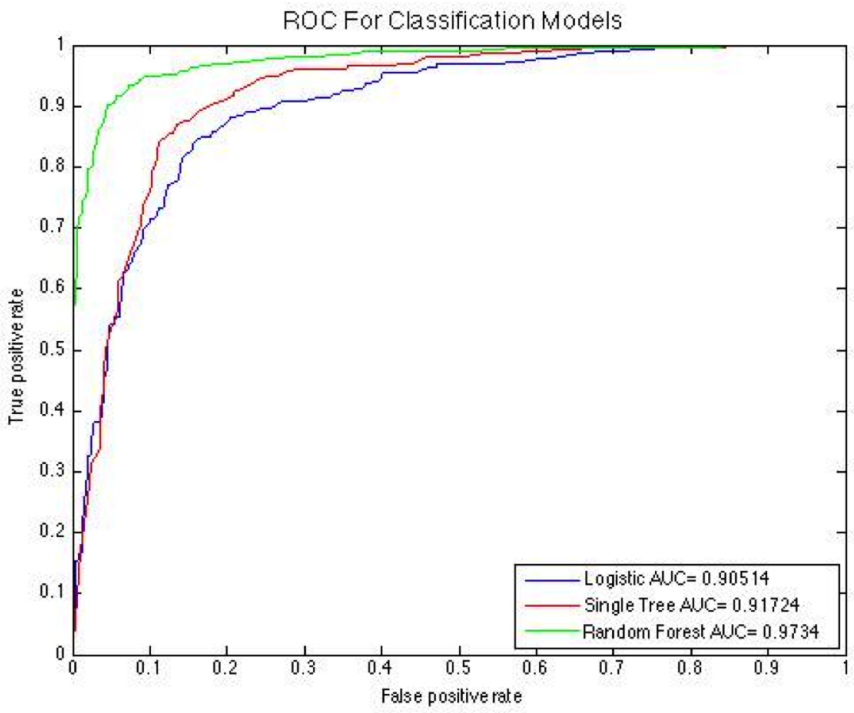


Figure 5-5: ROC curve for different ORD-EWR departure delay classification models.

temporal variables are the most significant while OD pair delay variables are generally the least significant. We assess the performance of the regression model by calculating both the mean and median of the error values for the 1,000 observations in the test set. For the 5 linear regression models, the average mean error was 34.2 minutes and the average median error was 24.4 minutes.

The second regression model is a single regression tree. We grow the trees using the same process as classification tress and investigate pruning. Figure 5-6 indicates that, similar to classification, the regression tree performs best when no pruning is conducted. The average mean error for this model is 22.7 minutes and the average median error is 5.3 minutes.

The third regression model we consider is random forest to build an ensemble of regression trees. As displayed in figure 5-7, the random forest regression error does not significantly decrease once more than 30 trees are built. As with classification, for the remainder of this thesis, we choose to use random forests with 40 trees. The average mean error for the random forest regression model using 40 trees is 21.5

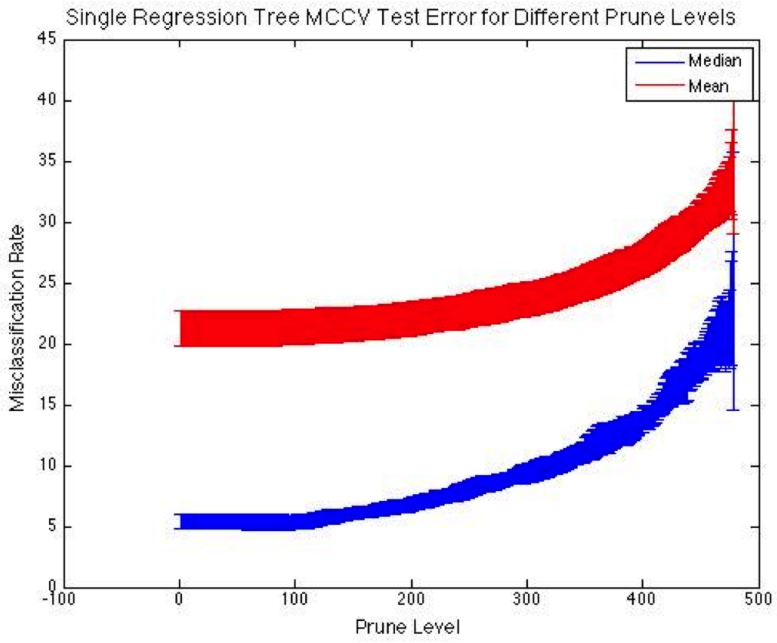


Figure 5-6: Single tree regression test error for different prune levels.

minutes and the average median error is 12.2 minutes.

A summary of all of the ORD-EWR delay prediction models is presented in Table 5.2. We use the random forest models for the rest of the thesis since they demonstrate the lowest mean error for both classification and regression.

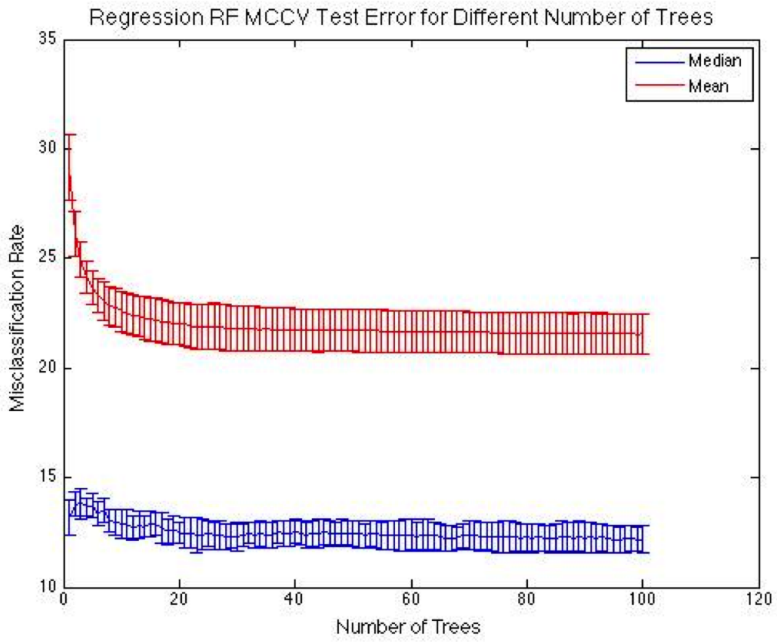


Figure 5-7: Random forest regression error for different numbers of trees.

MCCV Results: Zeros Removed		
Classification	**Mean Error**	**AUC**
Logistic	16.4%	0.91
Single Tree	13.8%	0.92
Random Forest	8.4%	0.97
Regression	**Avg. Mean Error (Min)**	**Avg. Median Error (Min)**
Linear	34.2	24.4
Single Tree	22.7	5.3
Random Forest	21.5	12.2

Table 5.2: ORD-EWR departure delay prediction models.

Chapter 6

Delay Prediction Models for the 100 Most-Delayed OD Pairs

In this chapter we use the random forest delay prediction models developed in Chapter 5 and apply them to the top 100 most-delayed OD pairs.

6.1 Determination of the 100 Most-Delayed OD Pairs

We determine the top 100 most-delayed OD pairs by calculating, for each OD pair, the mean of the 10,000-most delayed hours for that OD pair. Table 6.1 displays the top 20 most delayed OD pairs. It is noteworthy that SFO exists as an airport in the top 8 most delayed OD pairs; however, most of these routes are regional flights and the other airports in the top 8 links are smaller regional airports. Our MCCV model creates a unique delay prediction model for each of the 100 OD pairs with the highest mean values.

6.2 Performance of Delay Prediction Models

In this section, we study the performance of the classification and regression departure delay prediction models for the 100 selected links; a 2 hour prediction window and a 60 minute classification threshold are assumed. We define regression error, which

OD Pair	Mean of The Top 10,000 Hours (min)
SBASFO	37.6
MFRSFO	35.0
ACVSFO	33.9
SMFSFO	32.9
SFOSMF	29.8
MRYSFO	28.8
SFOSBA	28.6
SFOACV	28.4
LGAFLL	28.3
ORDIAH	26.4
ORDEWR	25.7
SFOMFR	25.6
MDWEWR	25.5
DTWEWR	25.4
EWRSTL	25.3
FLLEWR	24.9
EUGSFO	24.8
MCOEWR	24.8
STLEWR	24.4
EWRORD	24.2

Table 6.1: The top 20 most delayed OD pairs.

we define as the average (of the five random forest models) median (of the 1,000 test points) difference between the model's expected delay and the actual interpolated expected delay. The average (among the 100 OD pairs) classification error for this model is 8.43% and the average (among the 100 OD pairs) median regression error is 13.04 min.

Table 6.2 lists the correlation between classification test error for an OD pair and explanatory variable importance. The three temporal variables all demonstrate a positive correlation with test error, particularly the day of week and season variables. This indicates that the temporal variables tend to have a high performance when delays are hard to predict from the delay state of the different links or airports, and the best option then is to make a prediction based on historical data. The moderate correlation of the test error with the previous type-of-day variable (0.39) may indicate that when delays are hard to predict, the information on the previous day becomes more important. Finally, neither the airports' or the OD pairs' explanatory variables showed a strong correlation with the test error. However, it is interesting to look at the sign of the correlation coefficient on OD pair delay. This means that having a high OD pair importance is loosely associated with having a lower error, which suggests

that OD pair delays can be a good predictor of future delays on a link.

Explanatory Variable	Correlation with classification test error
Time of Day	0.23
Day of Week	0.5
Season	0.49
NAS State 2hr Prior	0.29
Previous TypeOfDay	0.39
Airports Delay (Max importance)	0.07
OD Pair Delay (Max importance)	-0.26

Table 6.2: Correlation between the test error and explanatory variables importance.

Comparing regression errors among different OD pairs is complicated by the differences in delay distributions of OD pairs. It is plausible to suspect that these differences play a large role in determining regression error. For example, we hypothesize that the greater the average delay of a link, the more error that link will possess. Figure 6-1 demonstrates that this is in fact not true; the OD pairs with the largest regression error generally have low average departure delays. It is also possible that the spread in the delay data for a given OD pair may affect regression errors. Because of the non-normal delay distributions, we use mean absolute deviation to express the spread of the departure delay distributions of each OD pair. Figure 6-2 shows that there is no clear relationship between the mean absolute deviation of the delays of OD pairs and their regression errors.

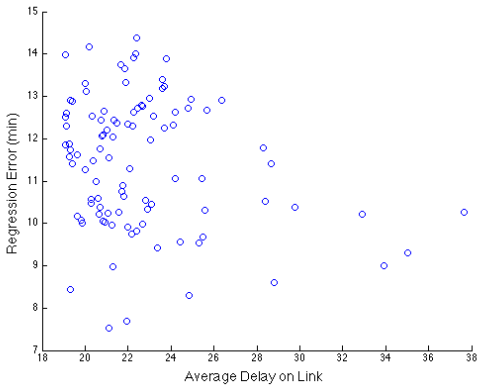


Figure 6-1: Regression error vs. average delay on the link

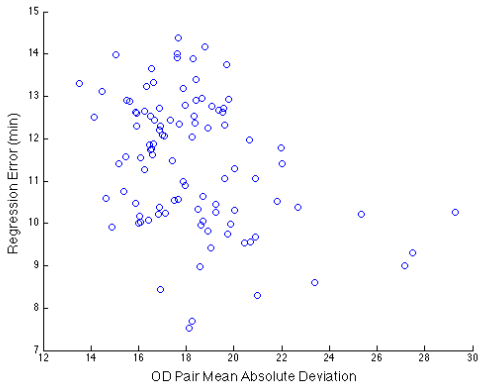


Figure 6-2: Regression error vs. mean absolute deviation on the link

6.3 Identification of the Most Influential OD Pairs

We try to identify which OD pairs' delay states play the most important role in predicting delays on the 100 most delayed OD pairs. To answer this question, we calculate the total importance value of each of the links' variables, which are obtained by summing the importance values of individual links' explanatory variables (both departure and arrival delay variables), over the 100 OD pairs. Table 6.3 shows the results of the analysis. The OD pairs at the top of Table 6.3 are those that better reflect the delay state of a certain area of the network, and consequently their associated delay variables play an important role in many of the links' delay prediction models. For example, the STL-ORD delay variable importance is over 70 for five different departure delay prediction models: STL-ORD, MEM-ORD, IND-ORD, SDF-ORD, and CLT-ORD.

Classification		Regresssion	
OD Pair	Total Importance	OD Pair	Total Importance
'SBASFO'	649	'SBASFO'	456
'ACVSFO'	566	'PDXSFO'	419
'PDXSFO'	563	'ACVSFO'	406
'MFRSFO'	554	'MFRSFO'	358
'PHXSFO'	477	'LASSFO'	309
'SNASFO'	444	'PHXSFO'	300
'LASSFO'	428	'SNASFO'	288
'RNOSFO'	405	'BOSEWR'	264
'BOSEWR'	403	'ORDEWR'	252
'MRYSFO'	388	'SANSFO'	250
'ORDEWR'	360	'RNOSFO'	246
'SMFSFO'	344	'DFWSFO'	243
'MDWEWR'	328	'SMFSFO'	242

Table 6.3: The most important links in the 100 most delayed OD pairs' prediction models.

We now choose three of the most influential OD pairs in Table 6.3 and take a closer look at the links for which they play an important prediction role. In Figures 6-3, 6-4 and 6-5, we see the links for which BOS-EWR, ORD-EWR, and PDX-SFO delay explanatory variables play an important prediction role. The colors of the links on the maps indicate the importance level with which each of the three selected links appears in the delay prediction models for the links that the arrows connect. All

three OD pairs display similar behavior in that all of the locations of the links for which the pairs play a role are all limited to a regional area; the east coast flights are unable to provide predictive capability of flights on the west coast and vice-versa.

Figures 6-6, 6-7 and 6-8 present aircraft rotation information for aircraft flying the route. For example, in Figure 6-6, the lines depict where aircraft flying from BOS to EWR flew into BOS from. For all three of our chosen OD pairs, the primary preceding route that is flown is the same exact same route, but in the opposite direction.

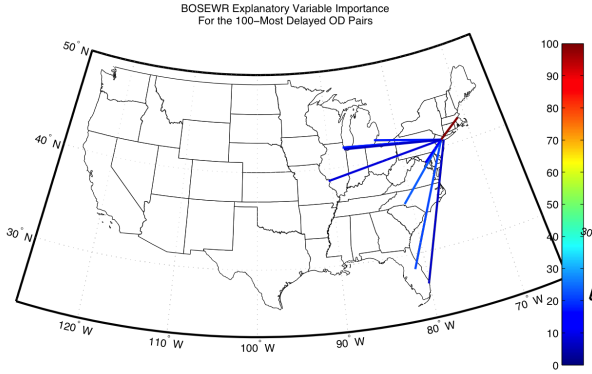


Figure 6-3: BOS-EWR explanatory variable importance for the 100 most delayed OD pairs.

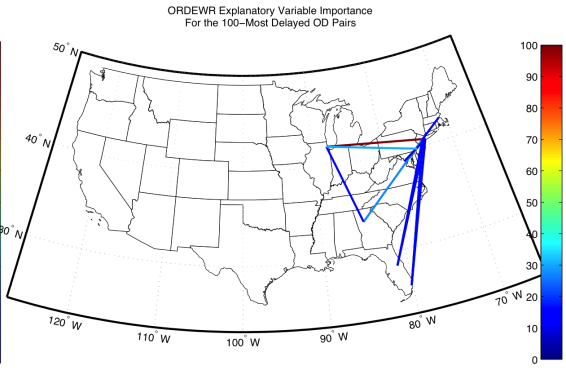


Figure 6-4: ORD-EWR explanatory variable importance for the 100 most delayed OD pairs.

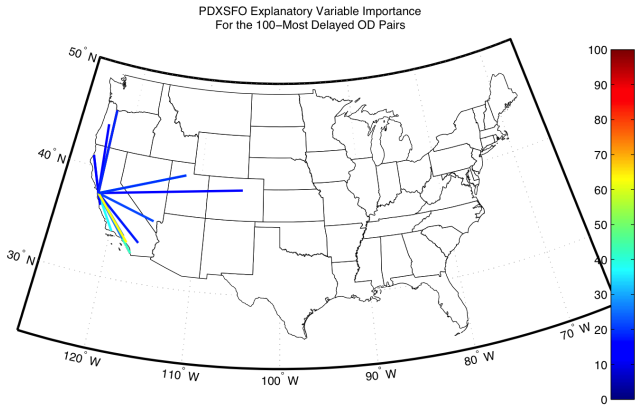


Figure 6-5: PDX-SFO explanatory variable importance for the 100 most delayed OD pairs.

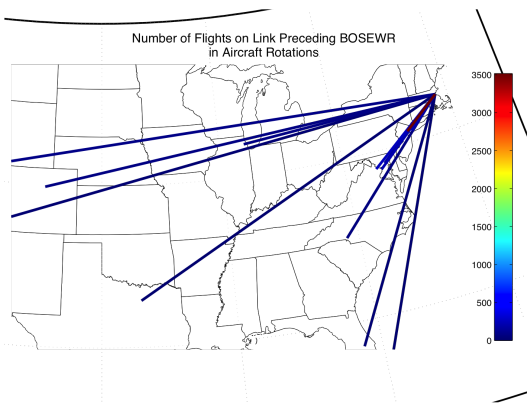


Figure 6-6: Number of flights on link preceding BOS-EWR in aircraft rotations.

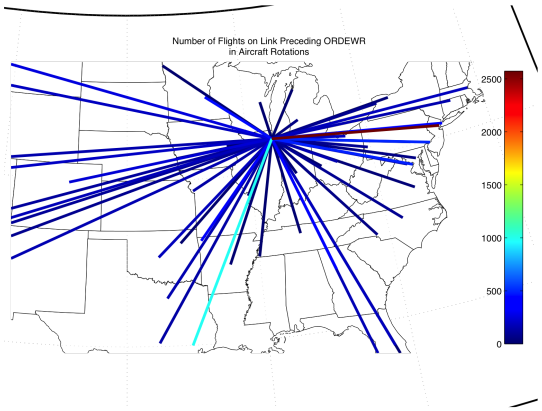


Figure 6-7: Number of flights on link preceding ORD-EWR in aircraft rotations

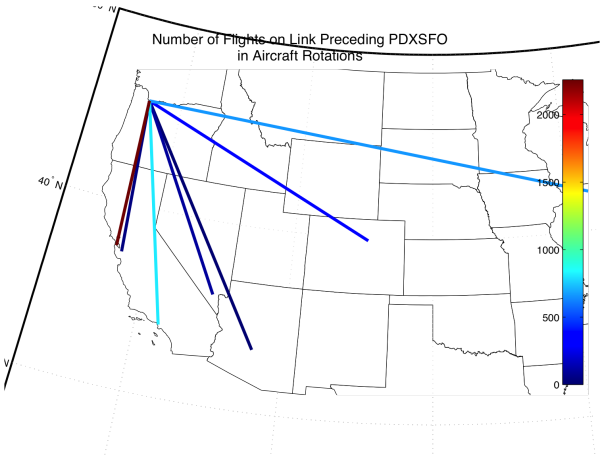
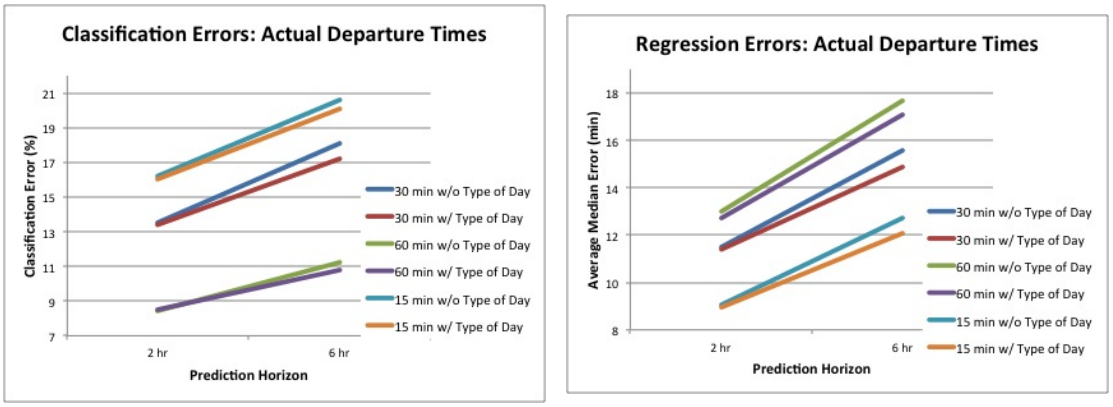


Figure 6-8: Number of flights on link preceding PDX-SFO in aircraft rotations

6.4 Effect of Changes in Classification Threshold

In this section we investigate what impact the classification threshold has on the delay prediction results. We continue to report the performance of models for the top 100 most-delayed OD pairs. We use actual gate departure and arrival times for this analysis.

Classification errors are lowest for a high classification threshold and classification errors are highest for a low threshold. This is reasonable, since very severe delays should be easier to detect. However, the opposite holds true for regression errors: low



(a) Classification

(b) Regression

Figure 6-9: MCCV errors at different thresholds and horizons using actual departure/arrival times

sampling thresholds yield the best accuracy while high sampling thresholds yield the worst accuracy. A graphical representation of these errors are shown in Figure 6-9.

As presented in Figure 6-10 the ratio of false positives to false negatives increases as we increase the classification threshold. Thus, at lower thresholds, more false negatives occur than false positives.

Figure 6-11 depicts the test error values for five different thresholds and the 100 links. The links are ordered according to their test error for a 60 min threshold. This plot shows that not all links have the same error reduction when the classification threshold is increased, and that this reduction is not correlated with the value of the test error. Figure 6-12 depicts the histogram of the test error increase when moving from a 90 min. threshold to a 45 min threshold. For most links the error increases by 5 percentage points(pp); but the increase ranges from as low as 1 pp to 10 pp.

Figure 6-12 provides a histogram of the test error increment for the top 100 OD pairs when the classification threshold is halved from 90 minutes to 45 minutes. We observe that the increment is not consistent among the OD pairs, which corroborates the results shown in Figure 6-11. Instead, the increments appear to be normally distributed, with an average error increment of 0.06 pp.

Figure 6-13 displays regression error histograms for the top 100 OD pairs at 30, 60, and 90 minute thresholds. Some outlying OD pairs are labelled. We observe that

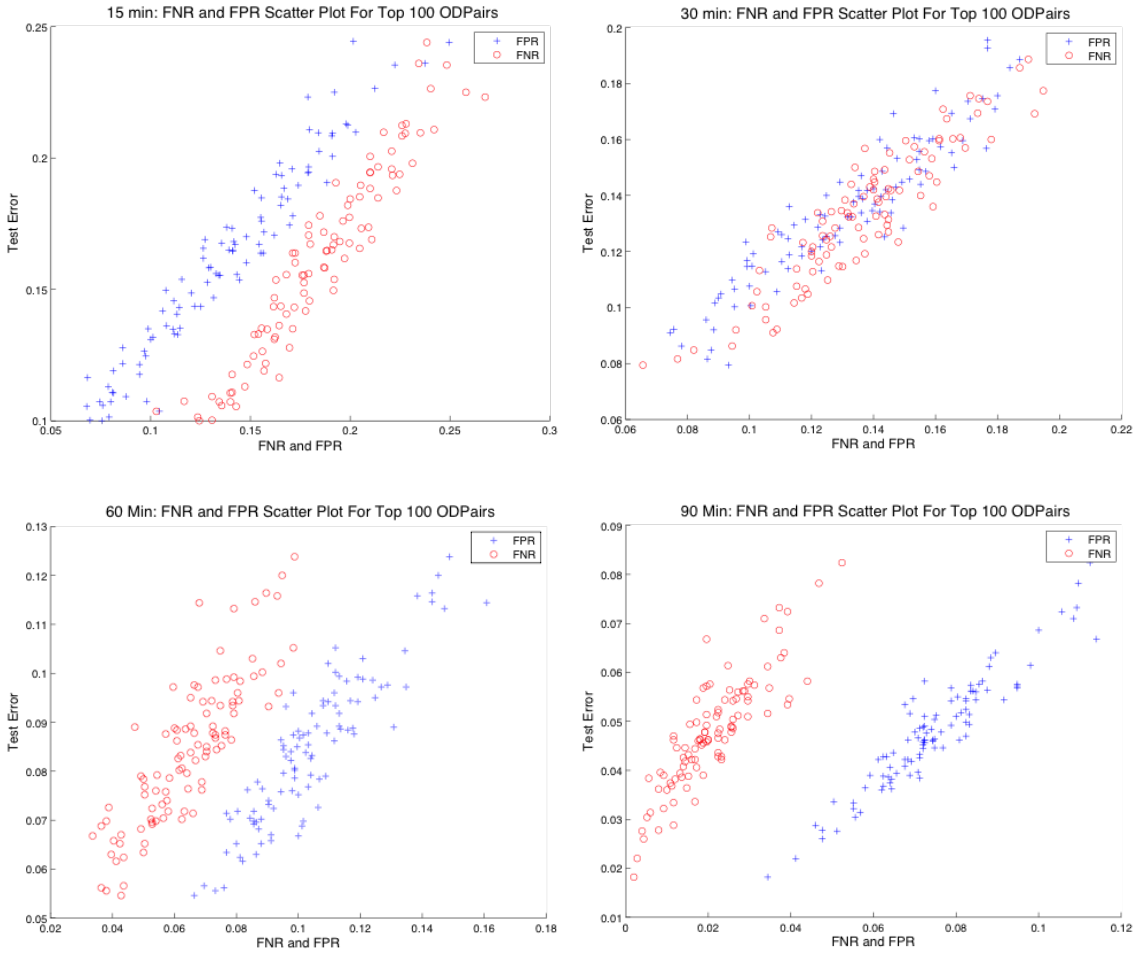


Figure 6-10: False positive rate and false negative rate for different classification thresholds.

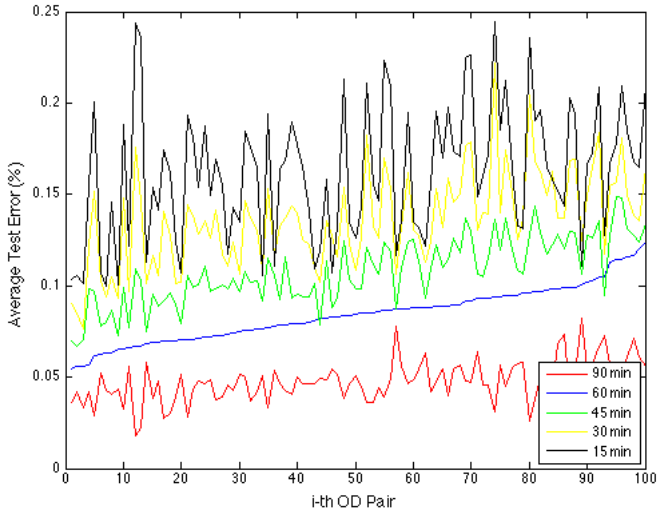


Figure 6-11: MCCV classification error for the top 100 most delayed OD pairs at different thresholds, ordered by 60 minute threshold error.

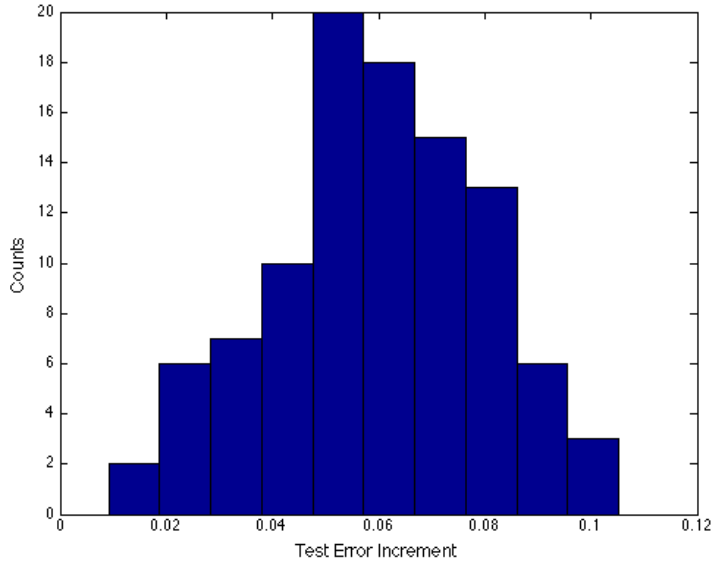


Figure 6-12: Histogram of the test error increment when changing the classification threshold from 90 min to 45 min.

some outlying OD pairs exhibit the largest regression error at multiple thresholds. Thus, even though we identified in Figures 6-11 and 6-12 that the change in test error as the threshold is changed is not strongly correlated with the magnitude of the test error, it still appears to have some correlation with outlying OD pairs. For example, XNA-ORD has one of the highest regression errors at all three classification thresholds.

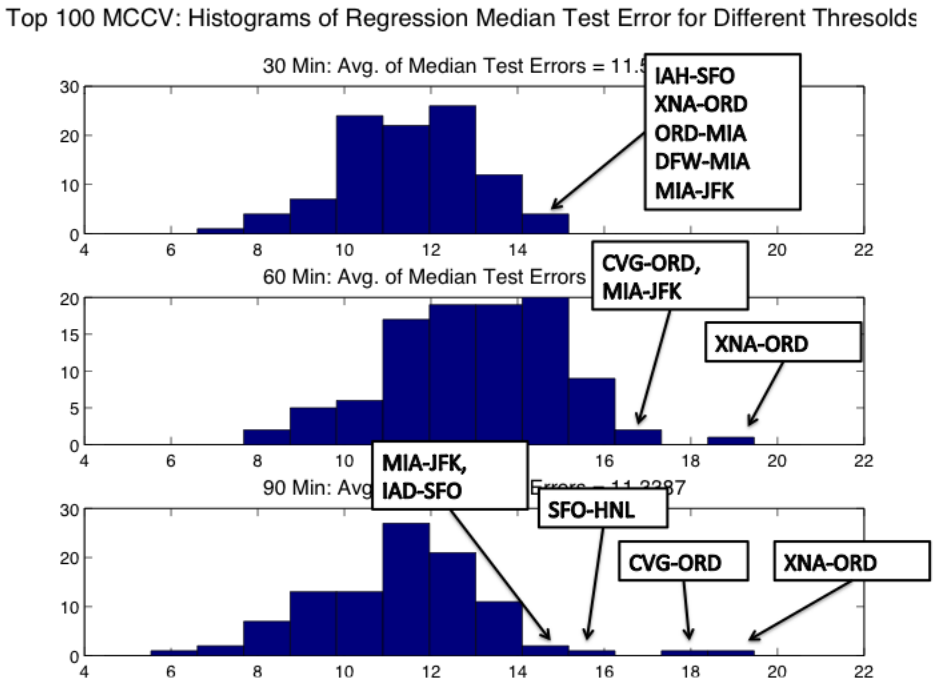


Figure 6-13: MCCV regression error histograms for the top 100 most delayed OD pairs at different thresholds. Some outlying OD pairs are labelled.

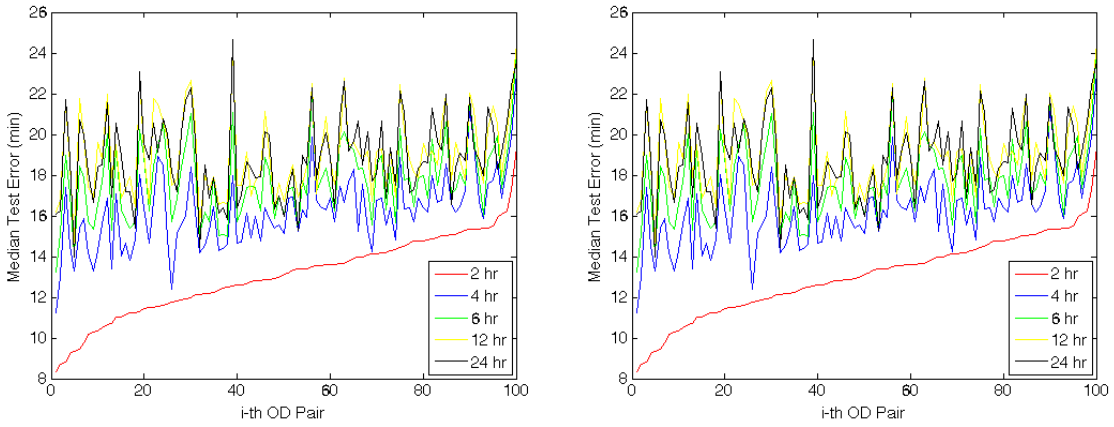
6.5 Effect of Changes in Prediction Horizon

In this section we investigate what impact the prediction horizon has on the delay prediction results. We continue to report the performance of models for the top 100 most-delayed OD pairs. We use actual gate departure and arrival times for this analysis.

For both classification and regression, as we increase the prediction horizon the prediction error increases. This is plausible since the states of the network elements

at the current time (NAS state, OD pair delays and airport delays) will become less and less relevant as we forecast further into the future. A graphical representation of the changes in prediction errors as we increase prediction horizon are shown in Figure 6-9.

Figure 6-14 shows the test error values for the 100 links arranged in increasing order according of the 2h horizon test error. There appears to be no correlation between the 2-hour horizon test error and the error increase as we increase the prediction horizon length.



(a) Classification (b) Regression

Figure 6-14: MCCV error for the top 100 most delayed OD pairs at different prediction horizons, ordered by 2-hour prediction horizon error.

Figure 6-15 plots regression error histograms for the top 100 OD pairs at 2, 6, and 12 hour horizons. Similar to Figure 6-13, some outlying OD pairs repeatedly demonstrate poor performance at all three horizons. XNA-ORD again appears to be one of the hardest OD pairs to predict at any given horizon. However, other outlying OD pairs, such as SBA-SFO, only perform much worse than the average at a specific horizon (in the case of SBA-SFO at a 12 hour horizon).

Table 6.4 shows the change in variable importance as horizon is increased for classification at a 30 minute threshold. The variable importance of the time of day increases as the prediction horizon increases. This is plausible, since the predictive capability of the delay states of other network elements decreases as the horizon is

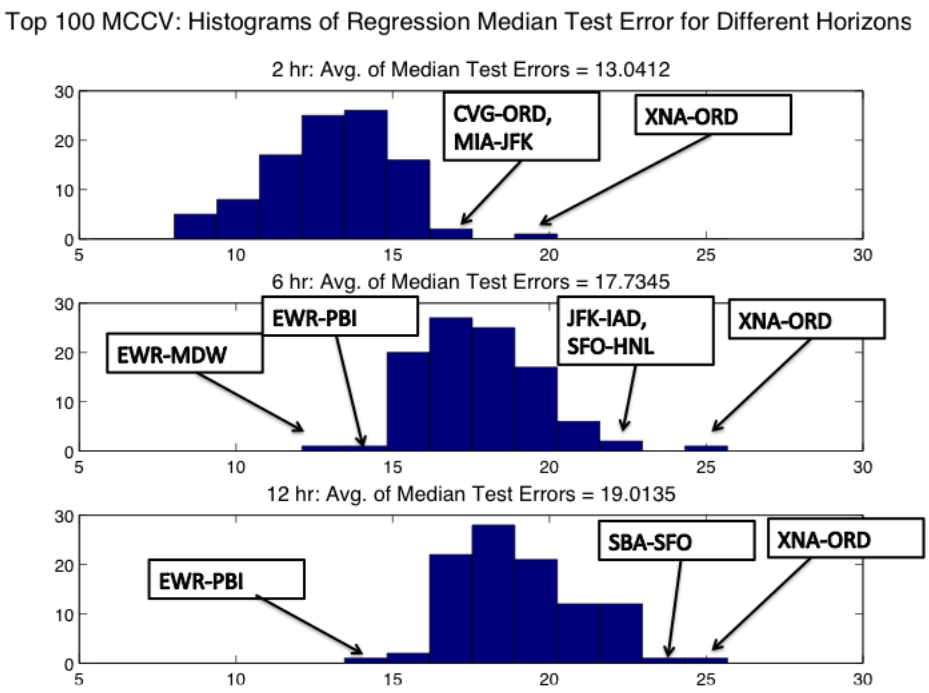


Figure 6-15: MCCV regression error histograms for the top 100 most delayed OD pairs at different horizons. Some outlying OD pairs are labelled.

increased, so in turn the time of day plays an increasingly important role in predictions. The importance of the previous type-of-day is noticeably low. It is noteworthy, however, that it is the highest around a 6 hour horizon. At a lower horizon, the specific OD pair and airport delays provide the most information for making predictions, and beyond a 6 hour horizon the behavior during the previous day begins to have less and less of an effect as more hours elapse in the current day.

	2 hr	4 hr	6 hr	12 hr	24 hr
Time of Day	73.2	98.9	99.8	100	100
Day of Week	22.9	38.1	40.8	37	34
Season	6.5	11.4	12.5	11.7	9.9
NAS Type of Hour	21.5	29.9	26.2	8.6	14.9
Previous Type of Day	13.6	23	25.2	24	16.3

Table 6.4: Average classification variable importance at different prediction horizons.

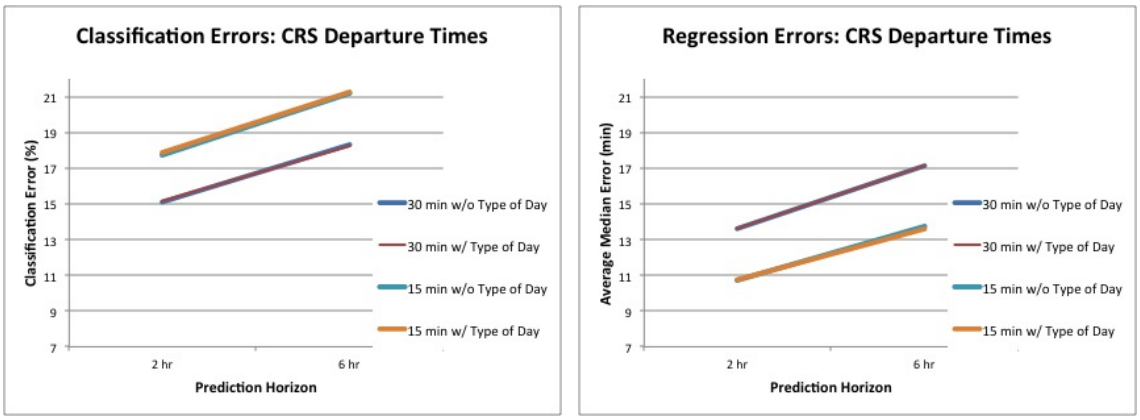
6.6 Using Scheduled Times

Previously in this thesis, all analysis has been done using actual gate times: the time when the aircraft actually departs and arrives at its gate in the airport. In this section we explore the effect of using scheduled gate times instead of actual gate times. Scheduled times are computer reservation system (CRS) times from the BTS database [7].

In order to use scheduled gate times, we re-bin and re-interpolate flights according to their scheduled times. This leads to new OD pair and airport delay variables. We establish new NAS state and NAS type-of-day variables using the same kmeans clustering process as before. Lastly, we create new type-of-day predictions following our previously discussed methodology.

We previously determined the top 100 most delayed OD pairs by calculating, for each OD pair, the mean of the top 10,000-most delayed hours for that OD pair. We used actual gate times for this calculation. Prior to running MCCV models on scheduled gate times, we first investigate whether the top 100 most delayed OD pairs are different when using the different dataset. Using the same metric, we determine that 79 of the original top 100 most delayed OD pairs are included when we instead use the scheduled gate times data. Thus, a majority of the top 100 most delayed OD pairs in the scheduled gate times dataset are also in the top 100 for the actual gate times dataset. We also find that all of the top 10 most delayed OD pairs for both datasets exist in the top 100 in the other. Despite these similarities, unless otherwise noted, when using scheduled gate times we choose to recalculate the top 100 most delayed OD pairs using the scheduled gate times delay data.

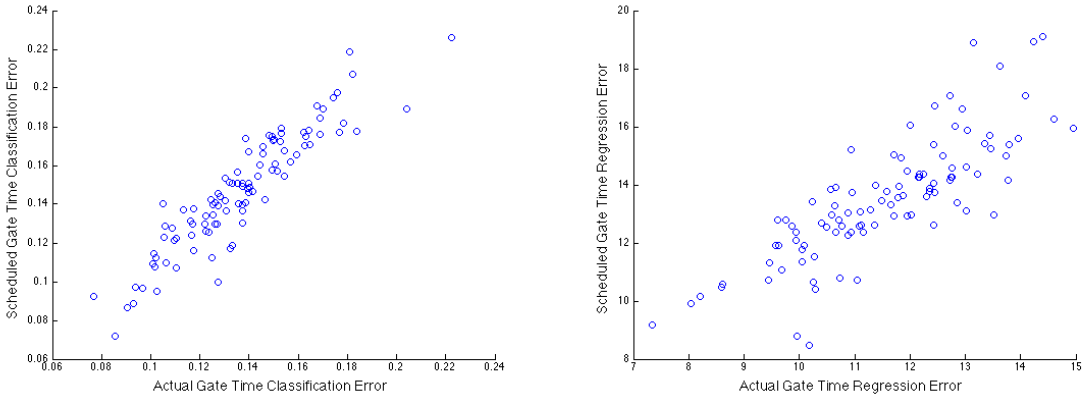
In general, the MCCV models using CRS times exhibit 1-2% more classification error and 1-2 min more regression error than the corresponding models using actual times. CRS and actual times exhibit similar increases in error when the prediction horizon is increased. A graphical representation of the CRS errors are shown in Figure 6-16. The errors resulting from using scheduled times at a 2-hour horizon most closely resemble using actual times at a 4-hour horizon.



(a) Classification (b) Regression

Figure 6-16: MCCV errors at different thresholds and horizons using CRS times

In Figure 6-17 we plot the MCCV errors for the top 100 most delayed OD pairs that result from using the actual gate times dataset and the CRS dataset. To accomplish this, we re-run the MCCV model using the CRS delay data, but we use the top 100 most delayed OD pairs from the actual gate times data. We use a 30 minute threshold and 2 hour horizon. The errors on a specific OD pair when gate times are used are very similar to the errors on the same OD pair when CRS times are used.



(a) Classification (b) Regression

Figure 6-17: MCCV errors for the top 100 most delayed OD pairs comparing actual gate times and scheduled gate times.

6.7 Using Wheel-Off Times

In this section we explore the effect of using wheel times, which correspond to the time when the aircraft leaves the ground on takeoff and the time the aircraft touches the ground upon landing. Wheel times will vary from gate times depending on the taxiing behavior of aircraft prior to and after each flight. However, the reported delay for each flight is still the difference between CRS time and actual gate time. There is no reported scheduled wheel-off time, so it is no possible to calculate delays based on wheel times.

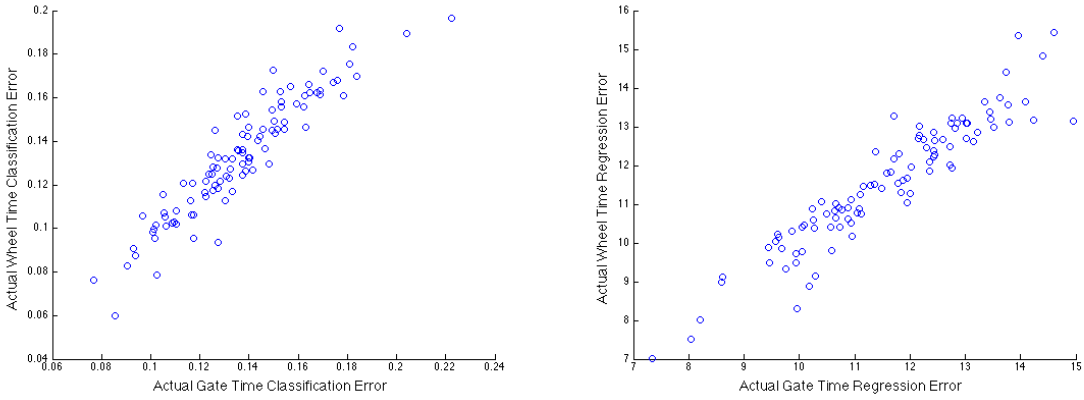
When using wheel-off times, we keep the NAS state and type-of-day k-means results that were obtained using departure delay gate-times. One area of possible further work is to re-do the k-means clusterings using wheel-off times.

We previously determined the top 100 most delayed OD pairs by calculating, for each OD pair, the mean of the top 10,000-most delayed hours for that OD pair. We used actual gate times for this calculation. Prior to running MCCV models on actual wheel-off times, we first investigate whether the top 100 most delayed OD pairs are different when using the different dataset. Using the same metric, we determine that 85 of the original top 100 most delayed OD pairs are included when we instead use the wheel-off times data. Thus, a majority of the top 100 most delayed OD pairs in the wheel-off times dataset are also in the top 100 for the actual gate times dataset. We also find that all of the top 10 most delayed OD pairs for both datasets exist in the top 100 in the other. Despite these similarities, unless otherwise noted, when using wheel-off times we choose to recalculate the top 100 most delayed OD pairs using the wheel-off times delay data.

MCCV errors when wheel-off times are used are very similar to MCCV errors using actual times. We focus our analysis on departure delay predictions at a 30 minute threshold. For a 30 minute threshold and 2 hour horizon, the classification and regression errors using wheel times are 13.4% and 11.6 min, respectively. The corresponding errors using actual times are 13.6% and 11.6 min. For a 30 minute threshold and 6 hour horizon, the errors using wheel times are 17.6% and 15.5 min,

which also closely resembles the errors for actual times: 18.1% and 15.6 min. This resemblance is plausible, since wheel times are very closely related to actual gate times. The only difference between the two lies in taxiing time, which generally accounts for only a small fraction of the gate-to-gate time of a flight. Figure 6-19 displays this behavior, as well as the effect of including NAS type-of-day, which is discussed in the next section.

In figure 6-18 we plot the MCCV errors for the top 100 most delayed OD pairs that result from using the actual gate times dataset and the actual wheel times dataset. To accomplish this, we re-run the MCCV model using the actual wheel times delay data, but we use the top 100 most delayed OD pairs from the actual gate times data. We use a 30 minute threshold and 2 hour horizon. The errors on a specific OD pair when gate times are used are very similar to the errors on the same OD pair when wheel times are used. Thus, we conclude that variability in taxi times does not significantly affect the model's performance.



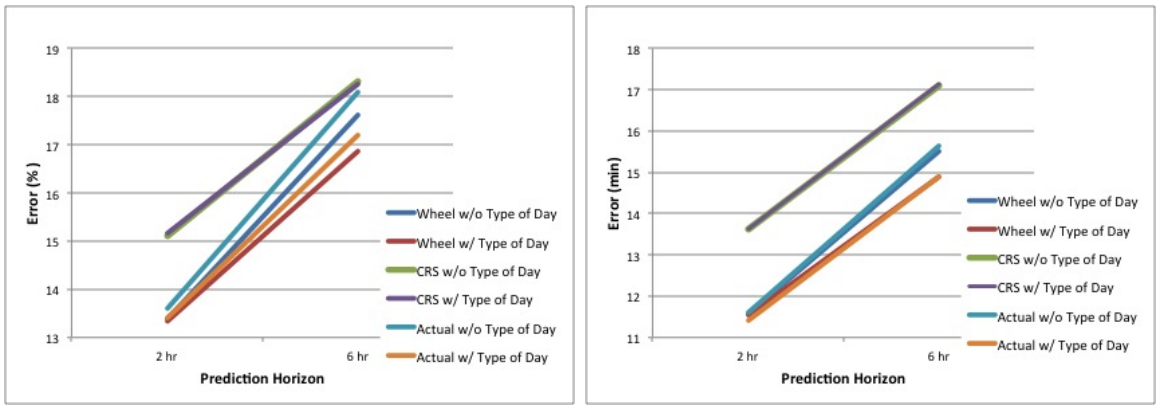
(a) Classification (b) Regression

Figure 6-18: MCCV errors for the top 100 most delayed OD pairs using actual gate times and actual wheel off/on times

6.8 Including NAS Type-of-Day in Delay Prediction Models

We include type-of-day as an additional explanatory variable in our delay prediction models. We choose to use the actual NAS type-of-day in the training set for the delay prediction models, but we introduce the predicted NAS type-of-day in the test set for the delay prediction models.

Interestingly, including NAS type-of-day improves models using actual times and wheel-off times, but it does not improve models using CRS times. Figure 6-19 displays this behavior. The models that exhibit the highest type-of-day and previous type-of-day variable importance are those that have a between a 4 and 12 hour prediction horizon and a low threshold (less than 45 min). Consequently, the errors in low threshold, large horizon models decrease the most when the type-of-day variable is added.



(a) Classification (b) Regression

Figure 6-19: Type-of-day effect on different datasets using a 30 minute threshold.

Figure 6-20 displays histograms of the regression error for the top 100 most delayed OD pairs using CRS times at a 30 minute threshold and 6 hour horizon. The top histogram corresponds to the model without type-of-day and the bottom histogram corresponds to the model with type-of-day included as an explanatory variable. The histograms are almost identical, corroborating our previous findings that including type-of-day does not have a significant effect on models using CRS times. We identify

some of the outlying OD pairs and conclude that the same OD pairs are outliers both in models with and without type-of-day.

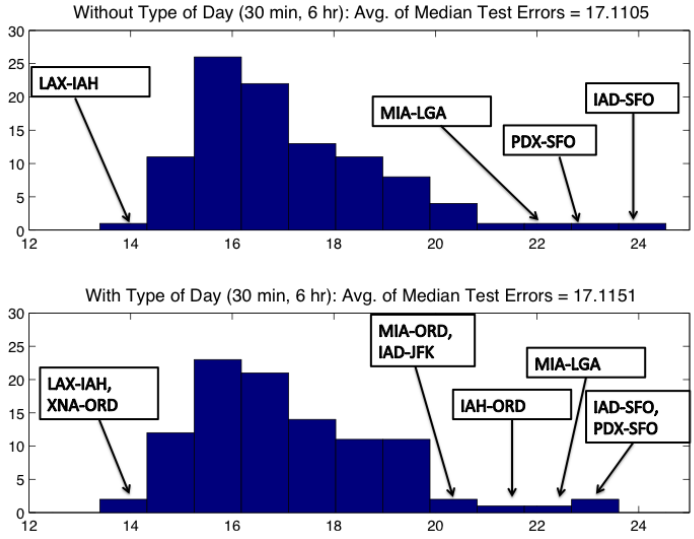
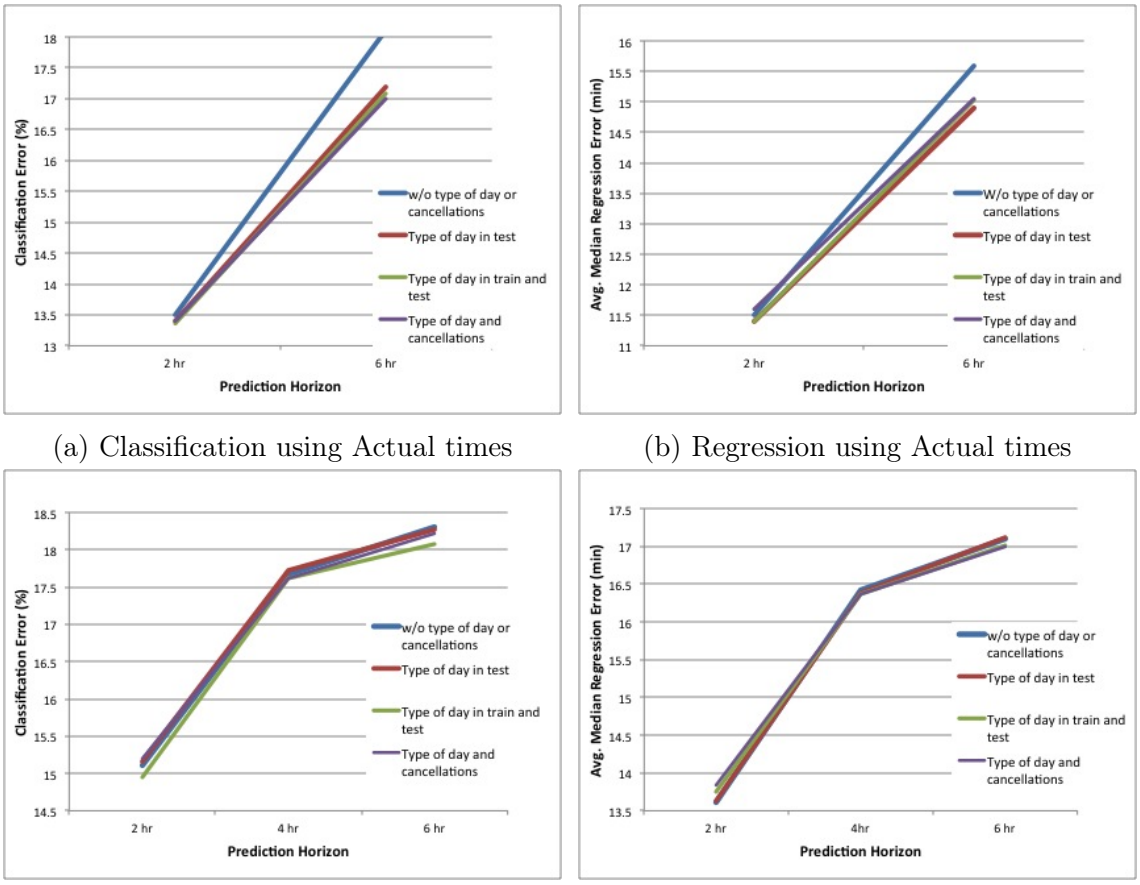


Figure 6-20: Regression error histograms for the top 100 most delayed OD pairs using CRS times at a 30 minute threshold and 6 hour horizon, with and without type-of-day included as a variable. Some outlying OD pairs are labelled.

Because type-of-day does not improve models using CRS times, we investigate whether including type-of-day in both the training and test sets yields a different result. As seen in Figure 6-21 this change decreases classification errors of CRS models but it does not significantly affect classification errors when using actual times nor does it affect regression errors. For the remainder of this thesis, when we include type-of-day in the models, we do so according to our first method; that is, we use the actual type-of-day when training and we use the predicted type-of-day when testing.

6.8.1 Error Assessment with Type-of-Day Included

In this section we revisit the performance of the delay prediction models by assessing their resultant errors when type-of-day is included. In particular, we aim to identify if there are any characteristics about a link which influences its ability to be predicted. We focus on cases using CRS times with a 30 minute threshold and with type-of-day included. Figure 6-22 displays multilayered plots of the regression errors vs. the average delay of the most delayed 10,000 hours for a given link (the same metric used



(a) Classification using Actual times

(b) Regression using Actual times

(c) Classification using CRS times

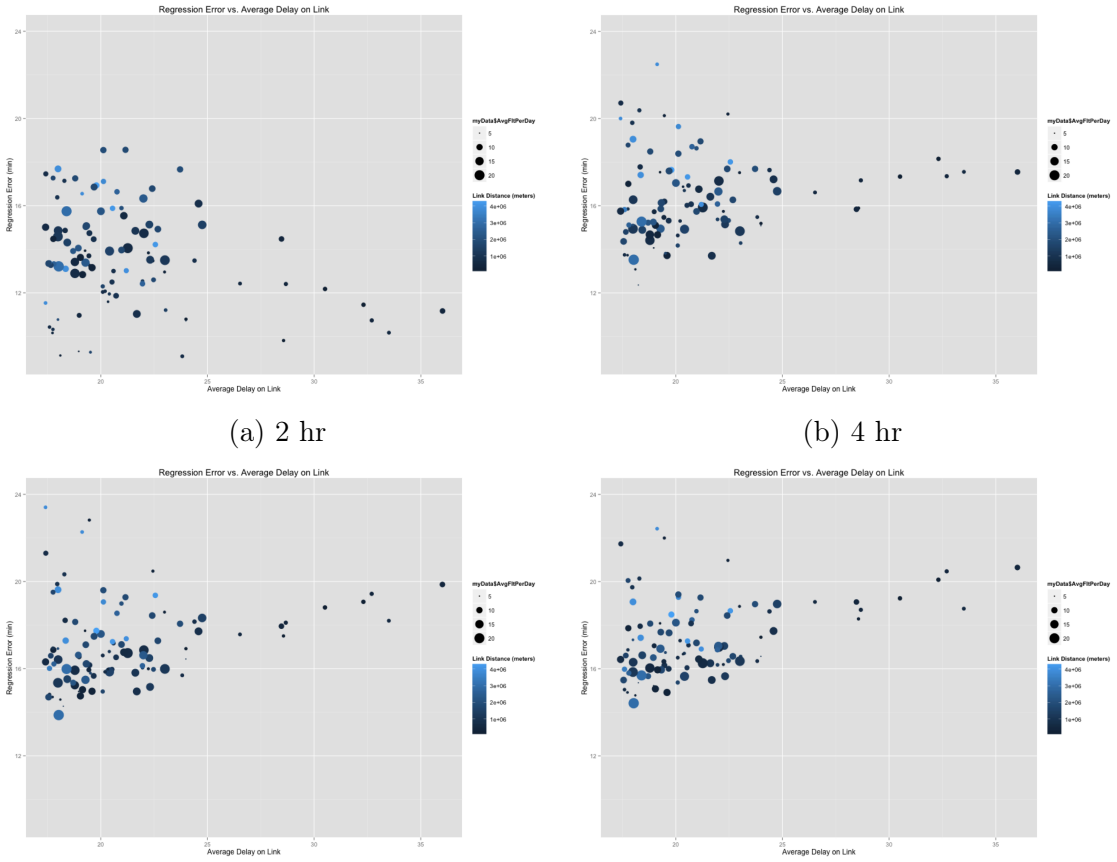
(d) Regression using CRS times

Figure 6-21: Type-of-Day effect on different MCCV models using a 30 minute threshold.

to determine the top 100 most delayed OD pairs). Average delay of the OD pair is graphed on the horizontal axis. Color corresponds to link distance: the lighter the color, the longer the link. The size of a point corresponds to flight frequency of the OD pair: the large the point, the greater the number of average flights per day on the link. At the 2 hour horizon, the OD pairs with high average delay are some of the easiest links to predict. However, as the horizon is increased, those OD pairs with high average delay exhibit a greater increase in error relative to OD pairs with low average delay. Thus, at a 12 hour horizon, OD pairs with high average delay are some of the hardest links to predict; the exact opposite of the 2 hour case. Figure 6-22 also demonstrates that the error on links with less frequent flights increases at a faster rate as prediction horizon is increased compared to error on links with more frequent flights. At the 2 hour horizon, the more frequently flown routes (represented by larger points) are some of the hardest flights to predict, but at the 12 hour horizon, they are the easiest flights to predict. We also observe two trends relating to link distance: the first is that nearly all routes with a greater than average delay on the link are short routes, while longer routes typically have less average delay. Second, at a 2 hour horizon, the longer cross-country flights are generally harder to predict, but as the horizon is increased this trend begins to reverse and shorter links become harder to predict. This trend is also observed in Figure 6-23, which maps the regression error of the top 100 most delayed OD pairs across the U.S. at 2, 4, and 6 hour horizons. Blue colors indicate low error while red colors indicate high error. As the horizon increases, the delays in cross-country flights generally become easier to predict relative to regional flights.

6.9 Comparison of OD Pairs With Best And Worst Performance

In this section we analyze and make comparisons between the OD pairs with the best and worst performance in the delay prediction models. The goal of this section is to



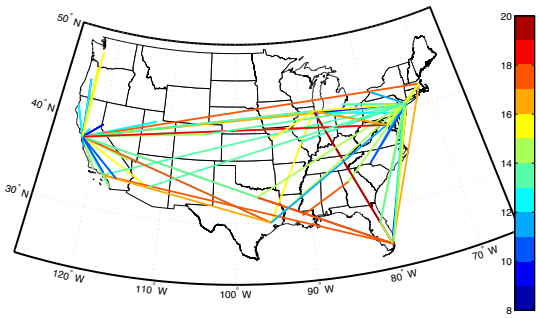
(a) 2 hr

(b) 4 hr

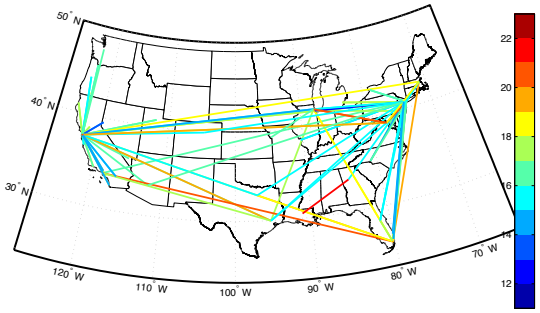
(c) 6 hr

(d) 12 hr

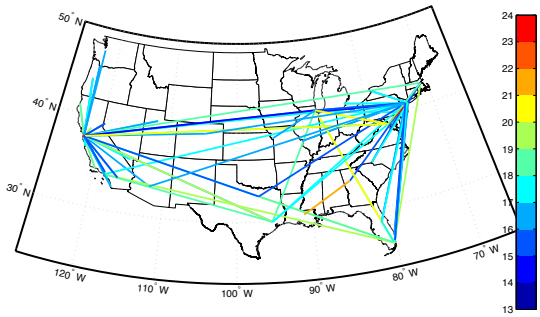
Figure 6-22: Regression errors of top 100 most delayed OD pairs at different horizons using CRS times and a 30 minute threshold.



(a) 2 hr



(b) 4 hr



(c) 6 hr

Figure 6-23: Map of regression errors of the top 100 most delayed OD pairs at different horizons using CRS times and a 30 minute threshold. To accentuate the differences in errors, the color scale is adjusted in each plot.

compare these two models, and understand what makes the prediction performance different among OD pairs. From the histograms in Figure 6-20, we see that LAX-IAH consistently has one of the lowest regression errors among the top 100 OD pairs, while IAD-SFO has one of the highest regression errors. We choose these two links for our analysis. At a 30 minute threshold and 6 hour horizon, IAD-SFO has the highest average median regression error among OD pairs (23.4 min) while LAX-IAH has the lowest error (13.9 min).

The time-of-day explanatory variable is the most important variable for both OD pairs. The difference in the models' performance can be explained using Figures 6-24 and 6-25. They show the IAD-SFO and LAX-IAH departure delay means and one standard deviation confidence intervals versus the time of day for the data points in the test set. We see that the IAD-SFO confidence intervals overlap less with the 30 minute threshold line than the LAX-IAH intervals. The more the overlap and lower the distance from the intervals' center to the 30 min threshold, the worse the prediction performance, because the difference between the likelihood of being above and below the decision threshold at a certain time decreases (we move towards a random guess). The IAD-SFO confidence intervals are also wider than the LAX-IAH intervals. This indicates lower correlation between the departure delay and the time-of-day variable, and an increased overlap with the threshold line.

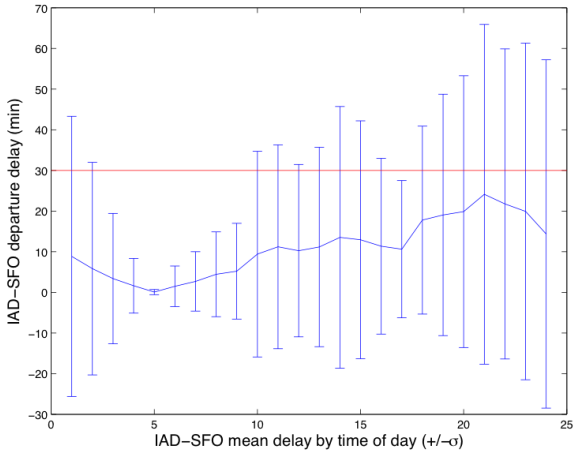


Figure 6-24: IAD-SFO mean delay by time of day ($\pm\sigma$).

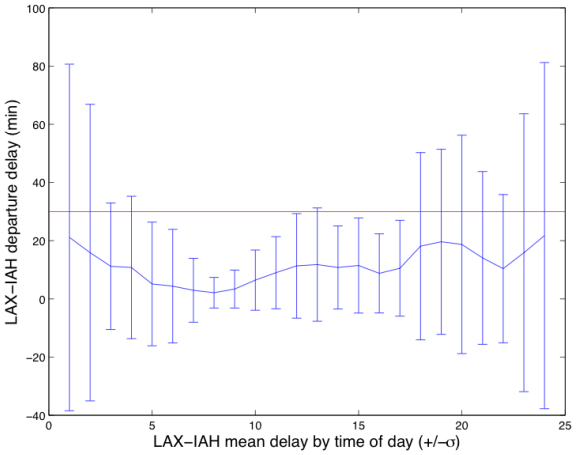


Figure 6-25: LAX-IAH mean delay by time of day ($\pm\sigma$).

6.10 Including Cancellations

In this section we add cancellations variables to the MCCV model. We introduce the cancellations variables in the same way as the OD pair delay variables and airport delay variables. Thus, we run a preliminary random forest consisting of 6 trees on 1200 oversampled data points to determine the top 100 cancellations variables. We then run three random forests consisting of 15 trees each and average the results to determine the 10 most influential cancellations variables to be included when we run the delay prediction model.

We initially include cancellations as an addition to the original type-of-day model; that is, type-of-day is used in the training set and predicted type-of-day is used in the testing set. The results are included in Figure 6-21. Cancellations do not generally have a noticeable effect on the model. The most noticeable effect occurs for classification when actual times are used with a 30 minute threshold and 6 hour horizon; in this case the classification and regression errors using cancellations are 17.0% and 15.05 min. The classification error is smaller than the classification errors resulting from the baseline model (18.1%, 15.64 min) and the type-of-day model (17.2%, 14.88 min). However, for regression and for other prediction horizons, the effect of adding cancellations is not discernible. The variable importance levels of the cancellations variables corroborate these findings; the variable importance of a cancellation variable is rarely greater than 5, and in no case is it greater than 10.

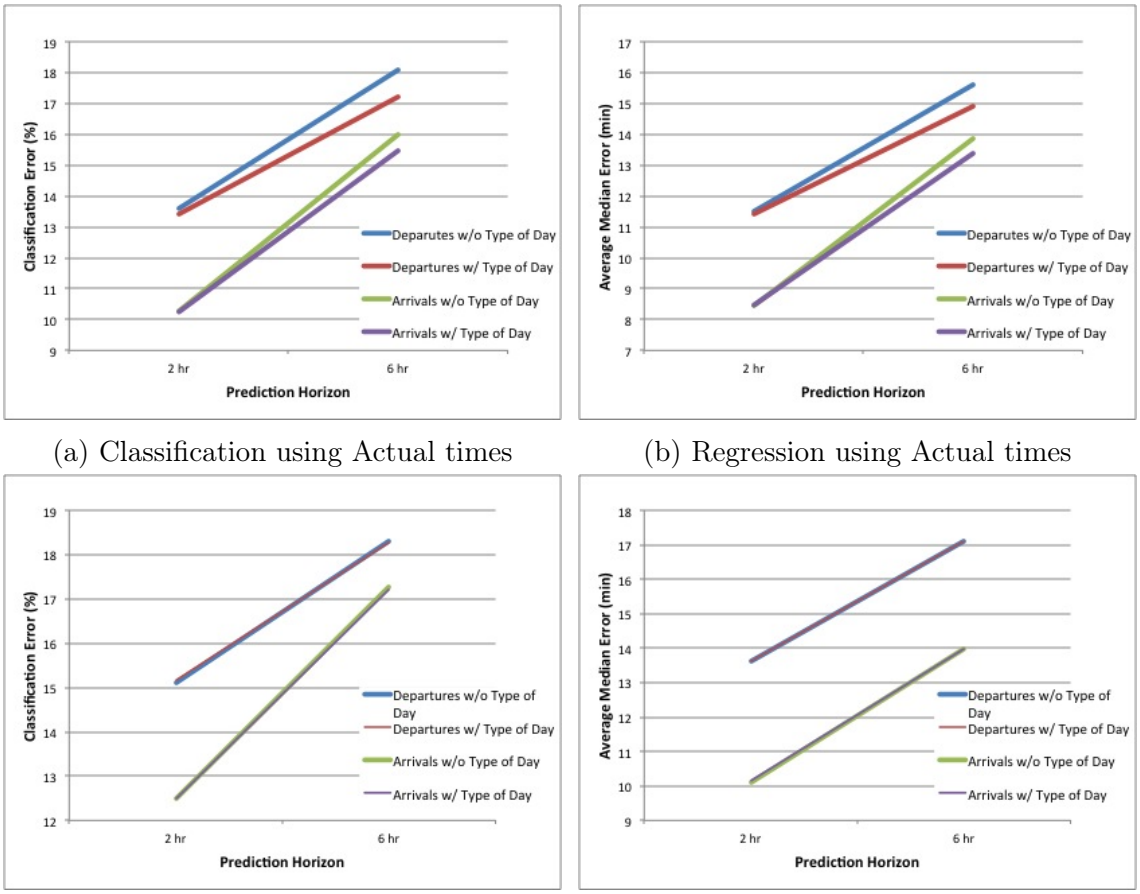
In future analysis, it may also be useful to see if cancellations and predicted type-of-day in both the training and test sets yields different results from those of the models already created.

6.11 Predicting Arrival delays

Thus far we have only predicted departure delays. We revise the previous models to predict arrival delays instead. The only change made to these models is the dependent variable (the expected delay for a specific OD pair in a given hour); all other variables

used in the random forest model remain unchanged. Note that this includes the NAS state and type-of-day variables. These variables were constructed using k-means clustering on departure delays. An area of possible further study is to reconstruct the NAS state and type-of-day variables using both departure and arrival delays.

As displayed in Figure 6-26, the arrival delay models behave similarly to the departure delay models. Both classification and regression errors increase as the prediction horizon increases. Adding NAS type-of-day to the model (when using actual times) decreases error by similar amounts as with departure delays; that is, NAS type-of-day has the greatest influence on the models around a 6 hour prediction horizon. Likewise, as before, adding NAS type-of-day does not have a convincingly noticeable effect on models using CRS times.



(a) Classification using Actual times

(b) Regression using Actual times

(c) Classification using CRS times

(d) Regression using CRS times

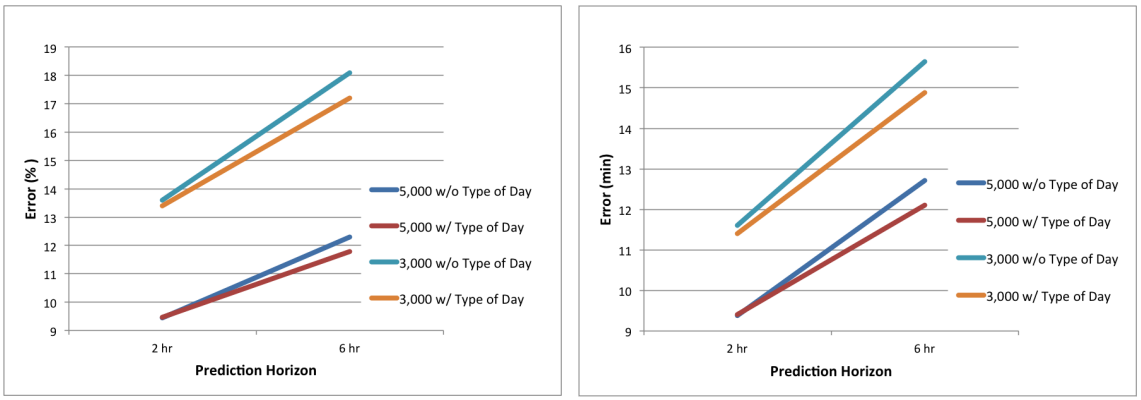
Figure 6-26: Arrival delay prediction using a 30 minute threshold.

6.12 Increasing MCCV Sample Size

In this section we explore the effect of the sample size of the random forest in the delay prediction models. Thus far we have used a training set consisting of 3,000 oversampled data points and a test set consisting of 1,000 oversampled points. Because there are 17,544 unique hours in the dataset, we have the ability to increase the sample size in the training set at the expense of computational efficiency. However, oversampling more than 3,000 points may also be problematic because certain OD pairs with low delays may not have 3,000 data points with a delay above the classification threshold. For example, JFK-SJU, the 100th most-delayed OD pair, only has 1,772 hours with delays greater than 30 minutes. Thus, oversampling to obtain a training set with more than 3,000 points will cause many points to be repeated multiple times in the training set.

We nevertheless run a delay prediction model using 5,000 points in the training set with a 30 minute threshold and 2 hour prediction horizon on actual times data. The model does not include type-of-day nor cancellations. The classification error for this model was 9.43% and 9.36 min. These errors are significantly lower than those from the same model which uses 3,000 points in the training set: 13.6% and 11.6 min. A similar disparity is seen for a 6 hour horizon: 5,000 points yields errors of 12.31% and 12.72 min while 3,000 points yields errors of 18.1% and 15.64 min. The same effect occurs with data that uses CRS times. The decrease in error as we increase the size of the training set is shown graphically in Figure 6-27.

We investigate whether including more data points in the training set decreases any variability in the classification and regression error of a specific run that is caused by the randomness of the selected training points. To explore this, we conduct 5 separate runs, using 5,000 training set points, of the top 100 MCCV model at a 30 minute threshold and 6 hour horizon using actual gate times. We choose not to include type-of-day. We then conduct an additional 5 runs using 3,000 training set points instead of 5,000. We observe that, in both the 3,000 and 5,000 point case, the variability among runs is very small. For both cases, the variability in error among



(a) Classification (b) Regression

Figure 6-27: Change in model error as training set size increases.

their 5 different respective runs was never greater than 0.1% for classification and 0.1 minutes for regression. The standard deviation in classification error was 4.5×10^{-4} for 3,000 points and 4.6×10^{-4} for 5,000 points. This indicates that 1) the variability in error among different runs is insignificantly small; and 2) the variability in error among different runs does not decrease as we increase the size of the training set. A similar conclusion can be reached for the variability in regression errors: 2.3×10^{-4} for 3,000 points and 7.1×10^{-4} for 5,000 points.

6.13 Effect of Oversampling in the Test Set

We also consider whether changing the oversampling method has an effect. We continue to use oversampling in the training set, but for testing we do not oversample. We do, however, still remove zero values as has been done previously. For a 30 min threshold and 2 hour horizon, the classification and regression errors are 15.6%, 18.2 min without the type-of-day variable. With type-of-day, these errors decrease to 15.3%, and 18.0 min. This decrease is more than the corresponding decrease when oversampling is used. Thus, oversampling could be masking some of the impact of using type-of-day.

6.14 Testing On 2013 Data

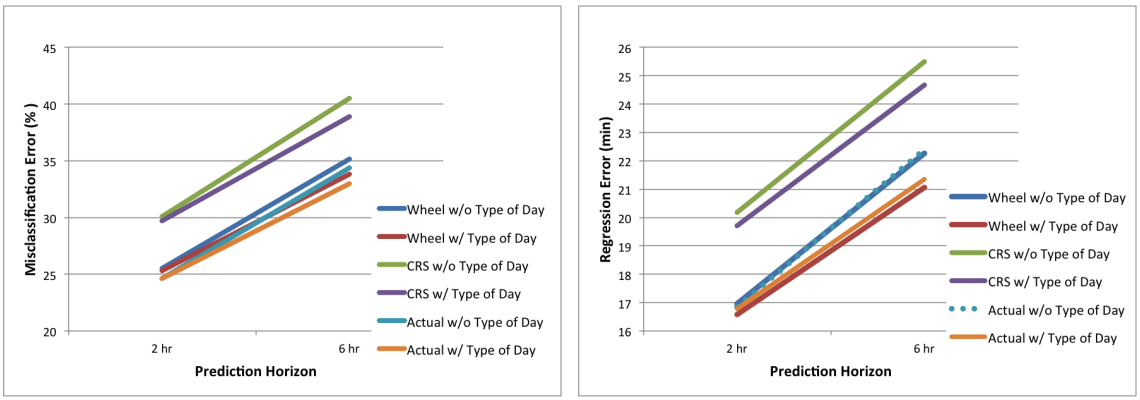
In this section we assess the performance of the delay prediction models using 2013 data as a second testing (validation) set. The 2013 data is preprocessed using the same methodology as the 2011-2012 data, described in Chapter 2. We continue to use the models that are trained on the 2011-2012 data.

6.14.1 Delay Prediction Model Performance

The classification and regression errors for MCCV models using actual gate times, a 30 minute threshold, a 2 hour horizon, and no type-of-day inclusion are 24.8%, 16.8 min, respectively. This is an increase in error from the previous test results that used data sampled from 2011-2012: 13.6% and 11.6 min. A similar behavior exists with the same model, but at a 6 hour horizon. The errors using the 2013 data for testing are 34.4% and 22.4 min. These error values are significantly higher than those obtained using the 2011-2012 data for testing: 18.1% and 15.6 min.

When we include NAS type-of-day in the same models, the resulting errors for a 2 hour horizon are 24.6% and 16.8 min. For the 6 hour horizon, the errors are 33.0% and 21.4 min. These error values are higher than the values obtained when testing using the 2011-2012 data. When type-of-day is included in the models tested on 2013 data, the addition causes a small decrease in classification and regression errors. This indicates that type-of-day has some predictive capability and positively influences the model. Figure 6-28 graphically summarizes these results.

As seen in Figure 6-28, when we use actual wheel on/off times instead of actual gate departure times, the delay prediction models behave similarly, which is to be expected give our previous results when testing on 2011-2012 data. When we use CRS times instead of actual gate times, there is a notable increase in both classification and regression errors, which is also similar to our results obtained when testing on 2011-2012 data. However, when testing on 2013 data, including type-of-day does have a significant effect on the CRS models. When 2011-2012 data was previously used to test the CRS models, type-of-day did not have a significant effect on classification or



(a) Classification (b) Regression

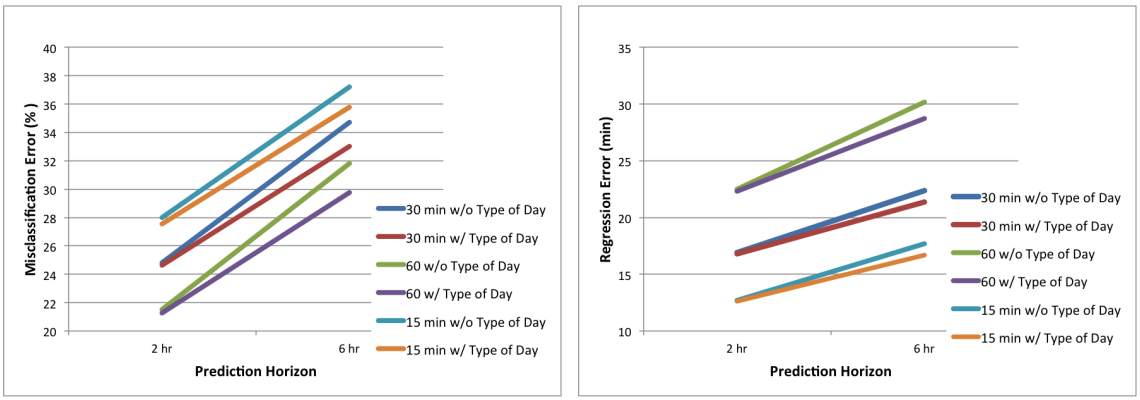
Figure 6-28: Delay prediction model 2013 test errors on different datasets using a 30 minute threshold.

regression.

To investigate whether the models tested on 2013 data are sensitive to the specified classification threshold, we re-run the same models at a 15 and 60 minute classification thresholds. In both cases, the results similar to those with a 30 minute threshold. For example, the errors on actual gate times at a 60 minute threshold and 6 hour horizon are 31.8% and 30.2 min. When type-of-day is included, these same errors are 29.8% and 28.7 min. Thus, even with a change in classification threshold, type-of-day continues to have a small positive effect on classification and regression models, although the magnitude of its effect does vary based on the threshold. Figure 6-29 displays these results.

6.15 2007-2008 Testing

In addition to comparing 2013 test errors to 2011-2012 test errors, we also use the 2007-2008 dataset for testing. As with testing on the 2013 dataset, when we test on the 2007-2008 dataset, we continue to use the 2011-2012 data for training the models. The 2007-2008 test errors are very similar to those obtained for the 2013 testing. For example, for actual gate times, a 30 minute threshold, a 2 hour horizon, without type-of-day included, the 2007-2008 errors are 24.7% and 17.2 min, which are similar to the 2013 errors: 24.8% and 16.9 min. When type-of-day is included, the 2007-2008



(a) Classification (b) Regression

Figure 6-29: Delay prediction model 2013 test errors on actual gate times and different thresholds.

errors are 24.4% and 17.0 min, which are also very similar to the 2013 errors: 24.6% and 16.8 min. This same behavior occurs at larger horizons and for the CRS times and wheel times datasets.

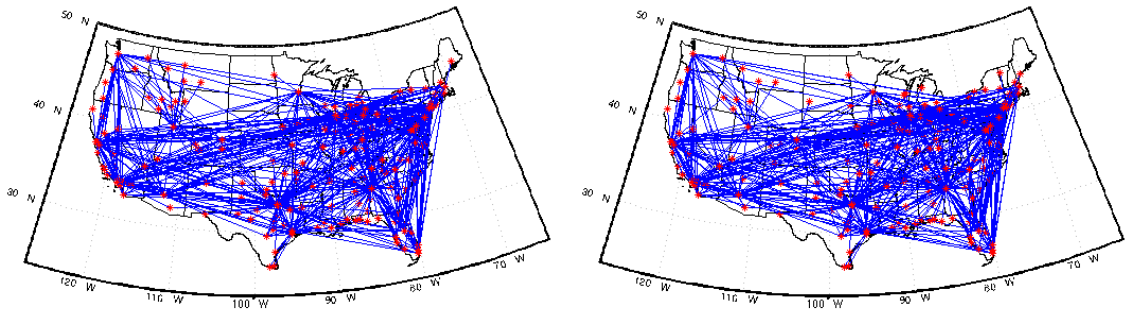
The similarity in test results between 2007-2008 and 2013 indicates that the delay prediction model's performance does not significantly vary year-to-year. Thus, even though the model is trained using 2011-2012 data, it is robust enough to use in different time frames.

Chapter 7

Temporal Analysis of NAS Behavior

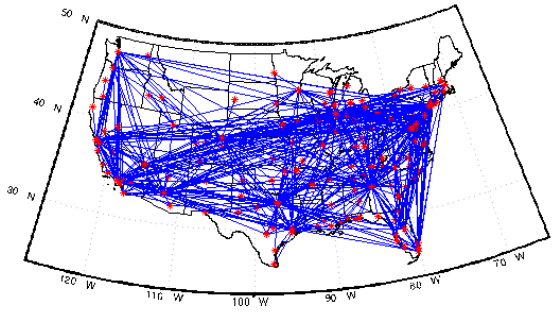
In this chapter we investigate changes in NAS Behavior over time. We do so by conducting the same k-means clustering processes for the NAS state and type-of-day variables for data from different years.

We use three different datasets for this purpose: 2003-2004, 2007-2008, and 2011-2012. The 2003-2004 data and 2007-2008 data is preprocessed using the same methodology as the 2011-2012 dataset, described in Chapter 2. We use the same 5 flights per day threshold for network simplification. This criterion creates slightly different simplified networks for the three datasets: there are 1281 unique OD pairs in the 2003-2004 simplified network, 1277 unique OD pairs in the 2007-2008 simplified network, and 1107 OD pairs in the 2011-2012 simplified network. This gradual decrease in the number of OD pairs with more than 5 flights per day average reflects a well-documented trend of consolidation of airline routes in the United States and a decrease in regional flights. Nevertheless, as displayed in Figure 7-1, the simplified networks for all three datasets are good representations of the NAS since they include OD pairs spread out across the entire United States and all major airline hubs are well represented.



(a) 2003-2004

(b) 2007-2008



(c) 2011-2012

Figure 7-1: Simplified Networks for different years.

7.1 NAS State Clustering

After network simplification and interpolation of delays according to the previously discussed methodology, we use k-means clustering to classify NAS states. Because we use 7 NAS state clusters for training the delay prediction models on the 2011-2012 dataset, we focus our analysis on k-means using 6, 7, and 8 clusters.

Table 7.1 lists the k-means clustering results for the 9 different configurations (three datasets with three numbers of clusters each). For each configuration, the New York City area, Chicago, and Atlanta all appear as unique high-delay clusters. I addition, each configuration has two clusters which do not have any particular delay centers; we consider these “Low NAS" clusters since the entire NAS is experience little to no delay.

Year	Number of Clusters	Low NAS	2nd Low NAS	NY	ORD	ATL	TX	2nd ORD/NY	SFO	WC	High NAS
	6	x	x	x	x	x	x				
2003-2004	7	x	x	x	x	x	x			x	
	8	x	x	x	x	x	x			x	x
	6	x	x	x	x	x		x			
2007-2008	7	x	x	x	x	x	x	x			
	8	x	x	x	x	x	x	x			x
	6	x	x	x	x	x			x		
2011-2012	7	x	x	x	x	x	x		x		
	8	x	x	x	x	x	x	x	x		

Table 7.1: NAS State k-means clustering results for different years and numbers of clusters.

The 6th, 7th, and 8th clusters vary by configuration. We notice that a cluster centroid representing high delays in Texas occurs in every configuration except for 6 clusters on the 07-08 and 11-12 datasets. Thus, we conclude that Texas is likely the fourth-most influential region in the NAS. The final two clusters vary more significantly by year. For 2003-2004, a “West Coast” cluster appears for 7 and 8 clusters, in which many airports along the West Coast experience moderate delays including some of the Southwest, such as PHX. One such cluster centroid is plotted in Figure 7-2. For 2007-2008, all configurations contain a second cluster in which both ORD and NY are delayed. Additionally, in two configurations, we see a “High NAS” cluster in which a large portion of the NAS is experiencing heavy delays. One such cluster centroid is plotted in Figure 7-3. For 2011-2012, as discussed in Section 3, SFO

occurs as being highly-delayed in one cluster. These differences indicate that there have been at slight changes to NAS delays in the last decade; however, since in all configurations at least 5 clusters are similar, we conclude that the majority of NAS delays have remained unchanged. Furthermore, we conclude the state of the NAS can be accurately represented by using between 6-8 clusters.

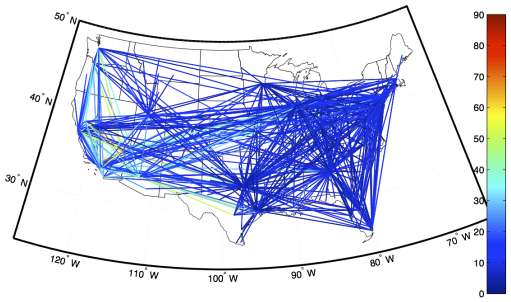


Figure 7-2: Example of West Coast cluster centroid.

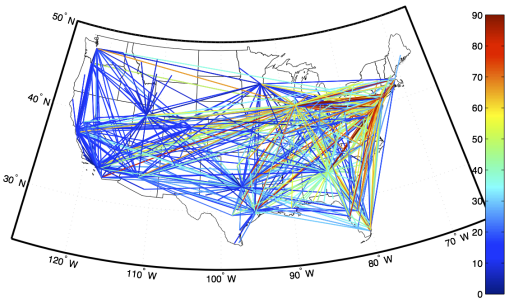


Figure 7-3: Example of high NAS cluster centroid.

7.2 NAS Type-of-Day Clustering

In this section we use the same k-means clustering process to classify the NAS type-of-day for the three datasets. In general the type-of-day cluster centroids exhibit more changes than the NAS state cluster centroids when we specify different years and different numbers of clusters. Unlike the NAS state clusters, which in general only contain one highly-delayed major airport hub or geographic area, the type-of-day clusters often contain multiple highly-delayed areas simultaneously, and to varying degrees.

Because we use 7 type-of-day clusters for training the delay prediction models on the 2011-2012 dataset, we focus our analysis on k-means using 6, 7, and 8 clusters.

The resulting clusters are evaluated based on each clusters' centroid delay values during each cluster's most heavily delayed hour in the day. Table 7.2 lists the k-means clustering results for the 9 different configurations (three datasets with three numbers of clusters each). In general, the type-of-day clusters are similar to the NAS state clusters. For each configuration, the New York City area, Chicago, and Atlanta all

appear as unique high-delay clusters. I addition, each configuration has one “Low NAS" cluster. With the exception of 6 clusters on 2011-2012 data, each configuration has a cluster in which the large Texas airports experience significant delays.

As with NAS state, the 6th, 7th, and 8th clusters vary by configuration. We observe that a second ORD/NY cluster appears for all but one of the 2003-2004 and 2007-2008 configurations. Conversely, the 2011-2012 configurations have a tendency to contain a 2nd low NAS cluster and a high SFO cluster instead. We also see that for every year group, a high NAS cluster exists for 7 and 8 clusters.

Year	Number of Clusters	Low NAS	2nd Low NAS	NY	ORD	ATL	TX	2nd ORD/NY	SFO	WC	High NAS
	6	x		x	x	x	x	x			
2003-2004	7	x		x	x	x	x	x			x
	8	x		x	x	x	x	x		x	x
	6	x		x	x	x	x				x
2007-2008	7	x		x	x	x	x	x			x
	8	x	x	x	x	x	x	x			x
	6	x	x	x	x	x			x		
2011-2012	7	x		x	x	x	x		x		x
	8	x	x	x	x	x	x		x		x

Table 7.2: Type-of-day k-means clustering results for different years and numbers of clusters.

We hypothesize that the frequent appearance of a highly-delayed SFO cluster, both for NAS state and type-of-day kmeans, could be due to construction on SFO runways that began in spring 2014. To test this hypothesis, we conduct a Wilcoxon rank sum test for difference in medians. The test yields a p-value of less than 0.0005, so we conclude the two distributions of the two years are different. We also separate the data from the two years and conduct k-means type-of-day clustering (k=7) independently on 2011 and 2012. In both years, a high-delay SFO cluster appears. A high ATL/NY cluster appears in 2011 that is replaced by a high NY/ORD cluster in 2012. This result indicates that while SFO delays in 2012 might be statistically different from delays in 2011, in both years SFO faces great enough delays to appear as a unique type-of-day cluster.

Chapter 8

NAS State Prediction

In this chapter we predict the NAS state at some hours in the future. We do so by using the NAS type-of-day predictions described in Section 4.2.1. For each type-of-day, we determine the sequence of NAS states that are closest to the NAS type-of-day cluster centroid. Thus, for each type-of-day, we have a sequence of 24 NAS states that represents the the most likely progression of NAS states throughout that specific type-of-day.

Table 8.1 shows these sequences. Each type-of-day begins and ends with at least one hour during which the most likely NAS state is the "Low NAS" state. For simplicity, Table 8.1 only documents the 'meaningful' NAS states for each day once the transition out of the "Low NAS" state has occurred. Figure 8-1 displays the same information in graphical format to show when the transitions from one NAS state to another are most likely to occur.

Type of Day	1st Meaningful NAS State	2nd NAS State	3rd NAS State	4th NAS State
1: High NAS	CHI	NYC	ATL	Med NAS
2: TX	Med NAS	TX	Med NAS	
3: Low NAS	Med NAS			
4: SFO	Med NAS	SFO		
5: ATL	Med NAS	ATL	Med NAS	
6: CHI	Med NAS	CHI	Med NAS	
7: NYC	Med NAS	NYC	Med NAS	

Table 8.1: Most likely NAS state progressions for each type-of-day.

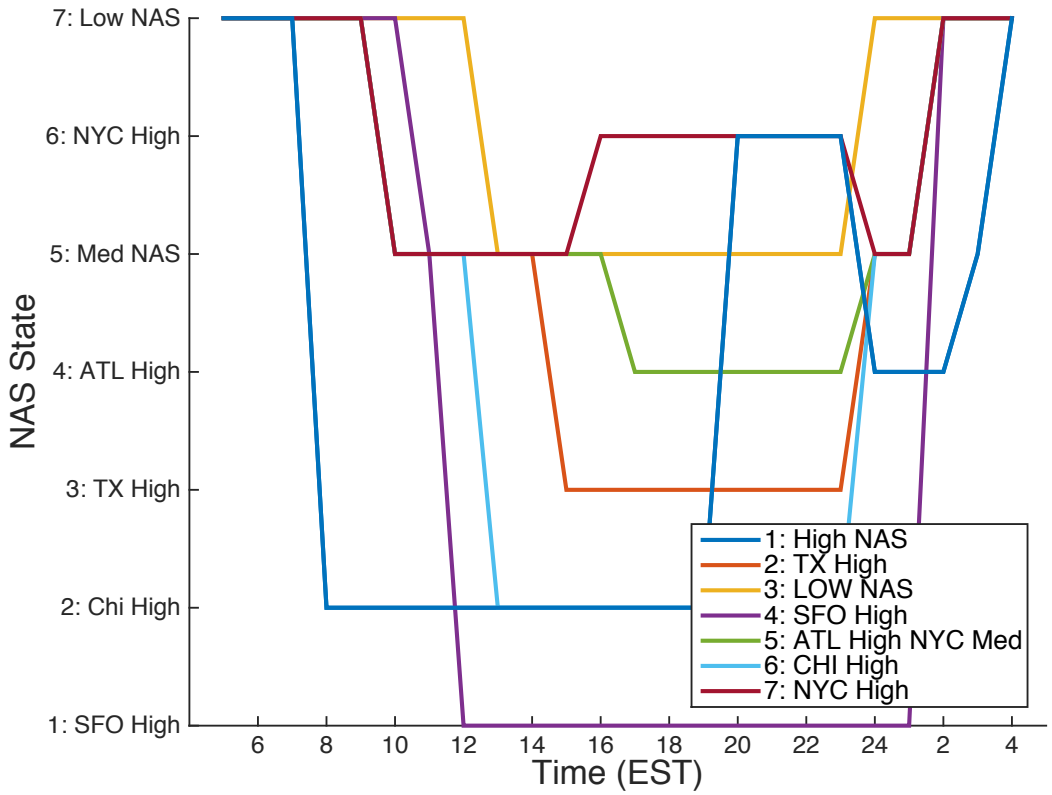


Figure 8-1: Most likely NAS state progressions for each type-of-day.

To predict the NAS state at h hours into the future, we use the NAS type-of-day prediction for the current hour t and find the NAS state that is closest to the NAS type-of-day cluster at $t + h$ hours. This process is akin to "following" the NAS state sequence for the predicted type-of-day for h hours. Figure 8-2 shows the prediction accuracy of this model at 2 hour and 6 hour horizons. The times listed are for the predicted hour (h+t). Thus the NAS state at 3:00pm EST is predicted with approximately 80% accuracy at both a 2 hour and 6 hour horizon (the actual prediction for the NAS state at 3:00pm would have been made at 1:00pm for a 2 hour horizon and at 9:00am for a 6 hour horizon).

Figure 8-2 shows that the model is robust to a change in prediction horizon. Prediction accuracy is only slightly worse at a 6-hour horizon than it is at a 2-hour horizon. Figure 8-2 indicates that the prediction accuracy is highly dependent on the time of day for which predictions are being made. At the start of the day, prediction accuracy is near 100% since every day begins at a "Low NAS" state. Prediction

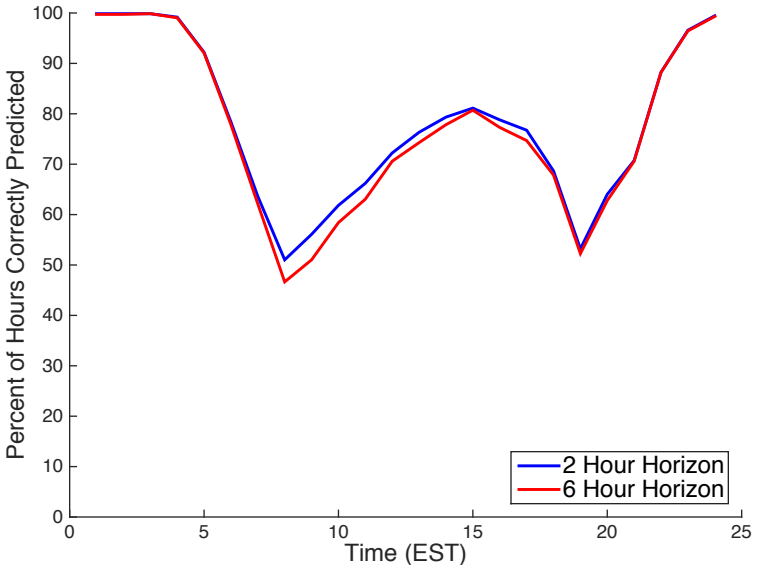


Figure 8-2: NAS state prediction accuracy.

accuracy decreases until it reaches a minimum of between 45-55% around 8-9:00 EST, which is when most days begin to transition to other NAS states. With so few hours in the day elapsed, the model has difficulty detecting which and when transitions will occur. As the day progresses, the model accuracy increases, since we usually observe repeated high NAS states for several hours during the middle part of the day. Prediction accuracy peaks at 3:00pm around 80%. As the NAS begins to transition back to "Low NAS" states, the model again demonstrates difficulty predicting when the transitions will occur, and consequently accuracy decreases. This is simply an error in timing; while the model understands the next transition will be to a lower delay NAS state, it does not know the exact hour it will occur. After the transitions to "Low NAS" states occur, the accuracy increases, as the model predicts, with greater and greater success, that the NAS is in a low-delay state as the day ends.

Chapter 9

Conclusion

This thesis presented methods for classifying and predicting U.S. airline delays. Our models used network-based delay statistics and thus focused on the effects of network-related delays and congestion in the NAS.

We preprocessed the data by simplifying the network to be studied, using only OD pairs that averaged at least 5 flights per day during 2011-2012. Individual flight data were then aggregated using a moving median filter with a 2 hour window a 1 hour time step. We interpolated delays when appropriate.

We identified several different types of explanatory variables to be included in the delay prediction models. In addition to temporal variables (time of day, day of week, season), we created OD pair delay variables that represented the expected delay on a certain route at a given time. Similarly, we created airport delay variables to represent the expected delay of a flight departing a specific airport at a given time.

We also used k-means clustering to create the NAS state variable, which gives a representation of the overall behavior of the U.S. airline network in a given hour. Similarly, we used k-means clustering to create the NAS type-of-day variable, which gives a representation of the overall behavior of the U.S. airline network over the course of an entire day. We performed these clustering techniques on three different time periods (2003-2004, 2007-2008, 2011-2012), which gave some insight into the changes in NAS delays over the last decade. For both NAS state and NAS type-of-day, the resultant clusters generally depicted delays that were centered around a

specific region or airline hub (NYC, Chicago, Atlanta, Texas, San Fransisco), which is a strong indication of network-based delays and congestion. We saw that the NAS type of day was seasonally dependent and that certain days were also linked to certain airlines. In order to properly use the NAS type-of-day variable in our models, we devised a method to predict the NAS type-of-day when only a small number of hours have passed in the day. We found that we can predict the NAS type-of-day at noon EST (using a test set) with 85% accuracy.

This thesis built upon formerly proposed delay prediction models by using the predicted NAS type-of-day in addition to the previous NAS type-of-day, NAS state, OD pair delay, airport delay, and temporal variables. Classification (logistic regression, classification tree, and random forest) and regression models (linear regression, regression tree, and random forest) were considered. For all models, data was oversampled to obtain a balanced dataset according to a specified threshold (15, 30, or 60 minutes). Random forest demonstrated the best testing performance for both classification and regression.

We created delay prediction models for the top 100 most delayed OD pairs. For each OD pair, a preliminary random forest was used for variable selection. Then, cross-validation was used to run random forest on 5 different training and test sets, from which the results were averaged. Including the NAS type-of-day variable benefited most models. We found that using actual departure times and a 30 minute threshold, average classification error for the top 100 OD pairs was less than 15% and average median regression error was less than 12 minutes for the 2011-2012 dataset. We investigated how certain properties changed the models' prediction accuracies, such as the oversampling/classification threshold, the prediction horizon, the length and flight volume of the route, and whether actual or scheduled gate departure and wheel off/on times were used. Lastly, we used these same models on new (2013) data, and we found that the models are relatively robust to more recent data.

Lastly, we used the NAS type-of-day variable to predict the NAS state at 2 and 6 hours into the future. We found that in general these predictions are relatively accurate; however, they suffer when predicting common state transition hours in the

late morning and in the evening.

The variables used in this paper, in particular the predicted NAS type-of-day, have the potential to be useful for making short-term delay predictions and for better understanding network-related behavior. While weather is indirectly incorporated in the models through the use of the OD pair delay, airport delay, and NAS state variables, future steps could be taken to more directly incorporate weather forecasts into these models so that they are more responsive to sudden changes in weather. Additional research could be undertaken to identify whether it is feasible to simplify the models by reducing the number of delay variables or by including regional delay variables instead. Other simplifications may exist that do not severely decrease model performance.

Bibliography

[1] Michael Ball et al. Total delay impact study: A comprehensive assessment of the costs and impacts of flight delay in the united states. Technical report, The FAA Consortium in Aviation Operations Research, December 2010.

[2] P. Fleurquin, J. Romasco, and V. Eguiluz. Data-driven modeling of systemic delay propagation under severe meteorological conditions. *arXiv preprint arXiv:1308.0438*, 2013.

[3] Pablo Fleurquin, José J. Ramasco, and Victor M. Eguiluz. Systemic delay propagation in the us airport network. *Sci. Rep.*, 3, 01 2013.

[4] A. Klein, C. Craun, and R.S. Lee. *Airport delay prediction using weather-impacted traffic index (WITI) model*. Digital Avionics Systems Conference (DASC), 2010.

[5] A. Klein et al. *Predicting Weather Impact on Air Traffic*. ICNS Conference, Herndon, VA, May 2007.

[6] K.B. Laskey, N. Xu, and C.H. Chen. Propagation of delays in the national airspace system. *arXiv preprint arXiv:1206.6859*, 2012.

[7] Bureau of Transporation Statistics. On-time performance transtats database.

[8] Bureau of Transporation Statistics. On-time performance- flight delays at a glance, 2014.

[9] Nikolas Pyrgiotis, Kerry M. Malone, and Amedeo Odoni. Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, 27(0):60 – 75, 2013. Selected papers from the Seventh Triennial Symposium on Transportation Analysis (TRISTAN VII).

[10] Juan Jose Rebollo. Characterization and prediction of air traffic delays. Master's thesis, Massachusetts Institute of Technology, May 2013.

[11] Y. Guan S. AhmadBeygi, A. Cohn and P. Belobaba. Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, 14(5):221–236, September 2008.

[12] Jinn-Tsai Wong and Shy-Chang Tsai. A survival model for flight delay propagation. *Journal of Air Transport Management*, 23(0):5 – 11, 2012.

[13] N. Xu, G. Donohue, K. B. Laskey, and C. H. Chen. *Estimation of Delay Propagation in the National Aviation System Using Bayesian Networks*. Proceedings of the 6th USA/Europe Air Traffic Management Research and Development Seminar, 2005.