Characterization and Prediction of Air Traffic Delays

by

Juan José Rebollo de la Bandera

Telecommunications Eng., University of Seville (2007)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author ....
Department of Aeronautics and Astronautics
May 17, 2012

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

Accepted by ....
Eytan H. Modiano
Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Characterization and Prediction of Air Traffic Delays

by

Juan José Rebollo de la Bandera

Submitted to the Department of Aeronautics and Astronautics
on May 17, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Aeronautics and Astronautics

Abstract

This thesis presents a new model for predicting delays in the National Airspace System (NAS). The proposed model uses Random Forest (RF) algorithms, considering both temporal and network delay states as explanatory variables. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and origin-destination (OD) pairs in the network, we propose new network delay variables that depict the global delay state of the entire NAS at the time of prediction. The local delay variables are identified by using a new methodology based on RF algorithms, and the importance levels of explanatory variables are used to select the most relevant variables. The high-level network delay variables are determined by using the k-means algorithm to cluster the delay state of different elements of the NAS.

The thesis analyzes both the classification and regression performance of the proposed prediction models, which are trained and validated on 2007 and 2008 ASPM data. The predictive capabilities of the models are evaluated on the 100 most delayed OD pairs in the NAS. The results show that given a 2-hour prediction horizon, the average test error across these 100 OD pairs is 19% when classifying delays as above or below 60 min. The study of the 100 most delayed OD pairs allows us to evaluate and compare prediction models for different OD pairs, and to identify models with similar characteristics. The effect of changes in the classification threshold and prediction horizon on model performance are also studied.

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

Acknowledgments

First, I would like to thank my advisor, Prof. Hamsa Balakrishnan, for her support and guidance through this work. She always provided exciting ideas, and help me to lead my research in the right direction.

I would like to thank my wife for making stressful moments easier and for her love and support. I would like to thank my parents, grandparents and brother for their unconditional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Flight delays in the US result in significant costs to airlines, passengers and society. In 2011, 20% percent of the flights were delayed more than 15 minutes [1]. Different studies have estimated the total cost of delays in the US [2, 3]. In 2007, the cost of domestic delays in the US economy was estimated to be $31.2 billion [2]. Such high delay costs show the need for better delay management mechanisms, and motivates the analysis and prediction of air traffic delays.

The large number of shared resources in the air traffic network, together with aircraft, crew and passenger interdependencies makes air traffic network effects an important field of study [4, 5]. Network effects are becoming more significant for two main reasons. Firstly, airlines attempt to increase aircraft utilization in order to increase their revenues, and thereby reduce the time buffers between arrivals and departures in their schedules. As a result, arrival delays become more likely to be propagated to subsequent departure flights [4]. Secondly, as demand approaches to capacity levels, the ability of the network to absorb disruptions decreases, making the system susceptible to large-scale delays. A study of network effects can help us understand factors that mitigate or amplify delay propagation, and to identify the elements of the network that have the most impact on the entire system.

The goal of this thesis is to study the potential of delay interdependencies in the National Aviation System (NAS) network in the development of delay prediction models. In particular, we are interested in predicting the departure delays of a particular OD pair by considering the current and/or past delay states of the different network elements. Similarly, we hypothesize that the delay state of different network elements at a certain time would be a good indicator of how NAS delays will evolve in the short term. We do, however, expect that our prediction models will have difficulties in capturing non-congestion related delays which only affect a few elements in the network (for example, delays related to mechanical issues which only affect a small subset of flights). It is important to note that our goal is not to predict individual flight delays, but instead to estimate future network-related delays on specific routes. We evaluate the prediction performance of this model over actual delay data, which includes all sources of delay.

Different delay prediction models have been proposed in the research community [5-10]. In [5] the authors study the propagation of delays in Europe, with the goal of identifying the main delay sources. In [6] a model for estimating flight departure delay distributions is developed, and the estimated delay information is used in an strategic departure delay prediction model. A different approach is presented in [7], the paper focuses on downstream delays caused by aircraft, cockpit and cabin crew. Other prediction models measure the impact of weather on delays, and they integrate weather information in delay prediction models [8-9]. In [10] a Bayesian network approach is proposed. The authors are able to capture interactions among airports by using a system-level Bayesian network. In contrast to these prior efforts, our models explicitly investigate the role of the network delay state in predicting delays.

We consider three different types of variables in our delay prediction models. First, we have temporal variables, which only depend on the time for which the prediction is being made (for example, the time of day or day of week), and not the delay state of the network. Second, we have local delay state variables, which indicate the delay level of specific elements of the network (for example the delay at a particular airport or on a given route). Finally, we have high-level delay state variables which depict the state of a group of elements, and are obtained by clustering local delay state variables.

A big challenge is the identification of the relevant delay state variables for each of the prediction models. We need to determine which routes, or airports delay states have a greater influence on the prediction model of interest. A new methodology to obtain these relevant variables is introduced in this thesis, using the explanatory variables' importance in a classification model to identify the relevant network elements.

The analysis of the different prediction models presented in this thesis will help us better understand delay interactions among the different elements of the NAS network, and evaluate how much of the future delay on a particular route can be explained by looking at the current network delay state. We evaluate two types of prediction models: classification models, where the output is a binary prediction that indicates whether the delay level is higher or lower than a predefined threshold, and regression models, where the continuous output directly estimates the delay along a route of interest.

1.1 Thesis organization

The rest of this thesis is organized as follows: Chapter 2 describes the data set used in this research and the preprocessing performed on it. Chapter 3 presents a high level analysis of NAS delays. The goal of this chapter is to identify the main delay patterns in the NAS. The results obtained in this chapter are used to generate categorical explanatory variables that describe the global delay state of the NAS. These variables are then included in the delay prediction models developed later in the thesis. In Chapter 4, the prediction performance of all the explanatory variables considered in this thesis is evaluated. Chapter 5 presents a comparative study of different regression and classification prediction models, and an exhaustive analysis of selected prediction models on the JFK-ORD route. Chapter 6 evaluates the performance of the prediction model chosen in Chapter 5 on the 100 OD pairs with the highest average delays in the NAS. The goal of this chapter is to validate the prediction model on different OD pairs, and to identify OD pairs with similar characteristics in terms of their prediction models' structures. Finally, conclusions and next steps of this research are discussed in Chapter 7.

Chapter 2

Dataset Overview and Preprocessing

In this chapter we describe the data set of study, and the filtering and aggregation performed on the data to enable the analyses presented in this thesis.

The results presented in this thesis were obtained using data from the FAA's Aviation System Performance Metrics (ASPM) database. The ASPM database integrates data from different sources: Enhanced Traffic Management System (ETMS), Aeronautical Radio, Inc. (ARINC), Official Airline Guide (OAG) and Airline Service Quality Performance (ASQP). ASPM provides detailed data for individual flights by phase of flight, along with airport weather data, runway configuration, and arrival and departure rates. Two years of ASPM data were used in our analysis, from January 2007 to December 2008. We processed the following ASPM fields for each flight:

- Dep_LOCID: Departure Location Identifier.
- Arr_LOCID: Arrival Location Identifier.
- SchInSec: Scheduled Gate-In Time.
- ActInSec: Actual Gate-In Time.
- SchOffSec: Scheduled Wheels-off Time.
- ActOffSec: Actual Wheels-off Time.
- FAACARRIER: Flight Carrier Code.

- TAILNO: Aircraft Tail Number.

These ASPM fields correspond to individual flight data; this data was then aggragated to obtain a more robust delay picture. The rest of this chapter describes the processing performed on the raw ASPM data.

2.1 Simplified network

The two years of ASPM data led to 2,029 airports, 31,905 origin destination pairs ("links"), and 22,795,187 flights. Our analysis aims at finding network effects, and correlations among OD pairs. We are therefore mainly interested in links with high traffic volume, which can have a significant impact on the rest of the network. Figure 2-1 depicts the histogram of the number of daily flights (on average) for each of the OD pairs. We can see that the majority of the links have fewer than one flight a day. Those links do not have enough traffic to impact the rest of the network significantly. For this reason, only OD pairs with 10 or more flights per day are included in the analysis in this thesis. Figure 2-2 shows the empirical cdf of the data from Figure 2-1, and we see that the 98th percentile of the links daily traffic distribution corresponds to about 10 flights/days. After the network simplification, the data set contains 532 links and 112 airports.

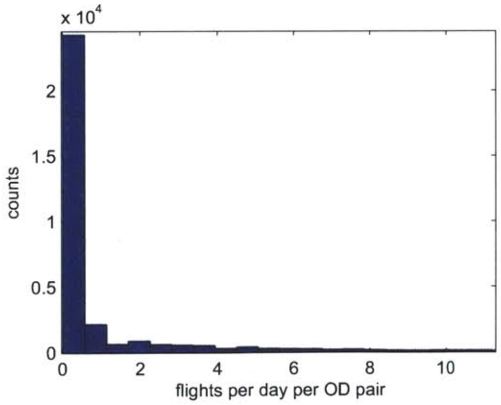


Figure 2-1: Histogram of the number of daily flights on different OD pairs.

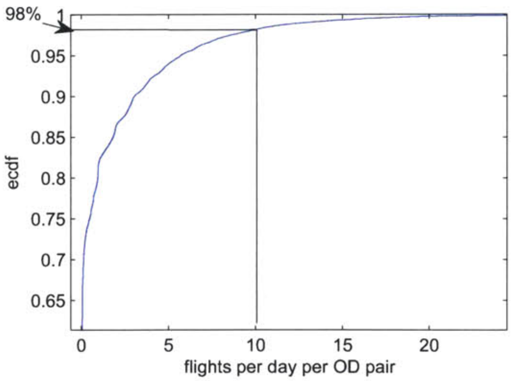


Figure 2-2: Empirical cdf of the number of daily flights on different OD pairs.

Figure 2-3 and Figure 2-4 show the OD pairs in the network before and after the network simplification; each link is represented with a blue arrow going from the OD central

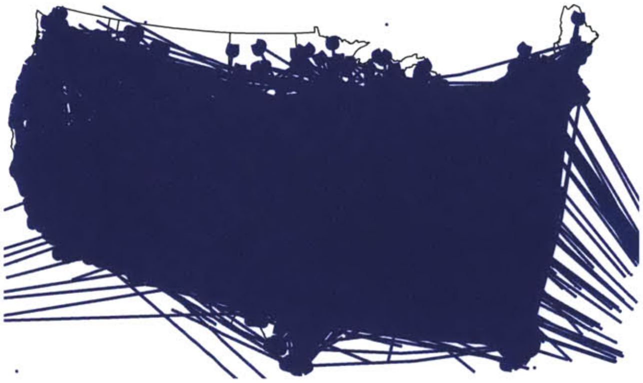


Figure 2-3: All links in the 2007-2008 ASPM data.

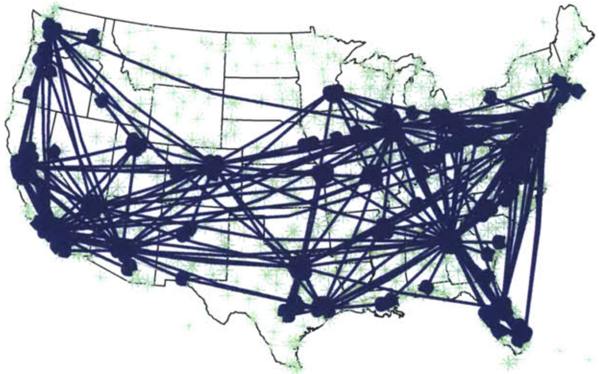


Figure 2-4: Simplified NAS network showing links with at least 10 flights a day. The light green icons denote airports in the original dataset that are not included in the simplified network.

point to the destination airport. We notice that the 532 selected links are fairly spread out, which will allow us to have a good picture of the state of the entire NAS. Table 2.1 presents a few metrics associated with the 20 airports with the highest number of arrival and departure links in the simplified network. The delay state of airports at the top of this list will potentially have a large influence on the overall NAS delay state due to their connectivity. ORD and ATL have the largest number of links, 90 and 82 respectively. None of the New York City area airports (LGA, EWR, JFK) is individually in the top five of this list, but if we add them together they have 74 links and they would be in third position. The last column of the table shows the connectivity within these 20 airports. This connectivity information indicates the potential impact of congestion in one of these airports on other high-traffic airports. ORD, ATL and DFW are the three airports with the highest level of network connectivity: 18, 15, 14 respectively. We note that these are also the airline hubs of United airlines, Delta Airlines and American Airlines, respectively.

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

Airport Id	Number of departure links	Number of arrival links	Connectivity among top 20
ORD	46	44	18
ATL	41	41	15
LAX	23	26	11
PHX	23	23	12
DEN	21	20	11
LAS	20	20	10
DFW	19	20	14
PHL	16	15	8
IAH	14	14	8
LGA	14	14	7
MCO	13	13	9
JFK	12	13	5
BOS	12	12	7
SEA	12	12	8
DTW	11	11	8
SFO	11	11	8
EWR	11	10	8
CTL	10	11	6
MSP	10	10	7
OAK	10	10	4

2.2 Individual flight delay data aggregation

The next step after identifying the simplified network is to aggregate individual flight delay data. We are interested not in predicting individual flight delays, but instead the delay levels of different airports and OD pairs in the network. We define the delay state of an airport or OD pair at time t as an estimate of the delay that a hypothetical flight using that resource at that time will experience. For example, if the BOS-MCO departure delay state is 30 min at 3 pm, it means that the estimated departure delay for a BOS-MCO flight taking off at 3pm is 30 min.

We use a moving median filter to obtain the delay states of airports and OD pairs. The delay state of any NAS element at time t refers to the median delay of all the flights that fall within a window of size W centered at time t. This low pass filter mitigates high frequency changes by calculating the median of the data points.

We chose a 1-hour step size for the moving window, which leads to 17,519 observations for the 2007-2008 data set. From time step to time step, independent of the window size, one new hour of traffic will be included in the window. We assume that the NAS delay state changes from hour to hour, allowing us to capture meaningful samples with a 1-hour time step. Figure 2-5 shows the histogram of the inter-flight separation times (between consecutive flights in the same link) in the simplified network. Sixty minutes corresponds to the 55th percentile of the distribution; 55% of the inter-flight separation times are under 60 min. More than half of the flights are separated less than 60 minutes, which supports the idea that the delay state of the links typically changes from hour to hour.

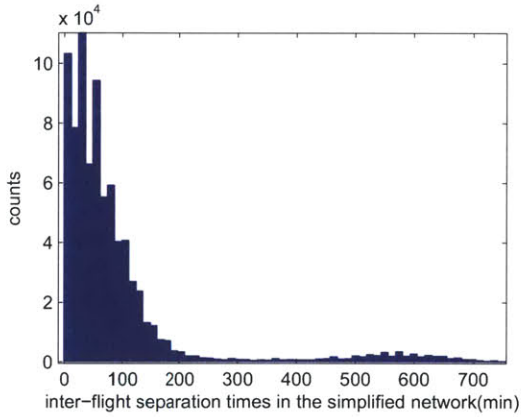


Figure 2-5: Histogram of the inter-flight separation times in the simplified network.

With respect to the window size, a narrow window will provide us with more information about how delay propagates in the NAS than a wide window. For example, with a 2-hour window we would need two consecutive observations to see how East Coast delays propagate to the West Coast, but with a 4-hour window we may only need one observation to see the effect on West Cost delays. At the same time, a wider time-window is more robust against outliers since information from more flights is accounted for when calculating the median delay. In the reminder of the thesis we focus on a 2-hour window, which shows good performance and also captures the average behavior without eliminating the NAS delay dynamics. We will also study the effect of changes on the window size on the results in Chapter 6.

An inherent problem of the proposed methodology is that we are estimating the delay state at time *t* (delay of an hypothetical flight using the network resource at time *t*) from the recorded delay data of flights falling in the corresponding time window. This raises the issue of how we might estimate the delay state value when there are no flights in the time window due to a lack of traffic. We could assume that the delay state is zero, but this will not always be true, especially when the delay state before and after that time period shows a high level of delay. We mitigate this problem by linearly interpolating the output of the moving median filter during periods with no flights in the time window, unless the period without flights lasted more than 6 hours, or the "end of day" (4 am) was included in the period without flights. In this manner, we did not interpolate between periods that were too separated in time, or between the end of a day and the beginning of the next one. Figure 2-6 shows the results of the interpolation methodology for a few days of data. In the circled area in Figure 2-6, we see that the original data was showing a delay increasing trend; however, due the lack of traffic the delay value was not always defined. On the other hand, the interpolation method fills the gaps in the data, giving us an smoother and more consistent delay curve.

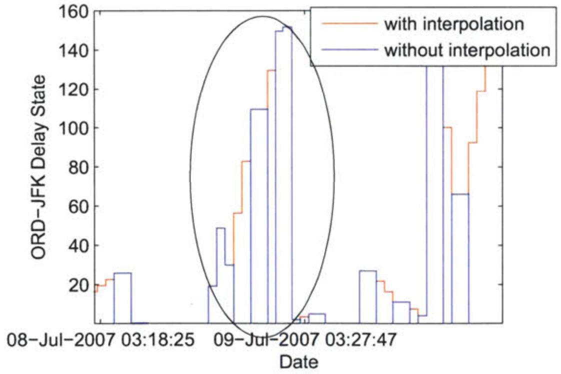


Figure 2-6: JFK-ORD departure delay state with and without interpolation.

After performing the described network simplification and data aggregation, the 2007-2008 data set leads to a network with 532 links and 112 airports, and 17,519 delay data points. There is on data point for each hour, and 2 hours of delay information is used to obtain each data point value.

Chapter 3

High Level Analysis of NAS Delays

In this chapter, the delay patterns of the entire NAS are studied. Our goal is to identify the typical delay states of the NAS. We believe that due to inter-dependencies among the different elements in the NAS and repetitive traffic patterns, we will be able to find a set of representative global NAS delay states.

To better understand the NAS temporal behavior, we will analyze the typical NAS states evolution with time; and will be able to answer questions like: When does delay typically spike, or drop? Which states occur more often? Which high delay states last longer? etc. We will also examine the composition of each of the typical NAS delay states; for example, we will identify airports and OD pairs exhibiting significant high delays, or interesting delay patterns.

In addition to spatial aggregation, we also explore temporal aggregation. In this chapter we obtain the typical NAS type of days by aggregating the NAS delay states for the entire day.

An important application of the typical NAS state identified is their use as explanatory variables in the delay prediction models that we propose later in the thesis. The global delay situation of the NAS has an important impact on the delay situation of specific OD pairs. We will see that the relevance of the categorical variables derived form the NAS states will depend on the characteristics of the OD pair for which we evaluate the prediction model.

3.1 NAS delay states

Our goal in this section is to identify the typical delay "snapshots" of the NAS. We cluster 17,519 NAS departure delay states (2007-2008 observations) into a few typical NAS delay states. The NAS delay state at time t is defined by the departure delay state of each link in the simplified network at time t. The typical NAS delay states are obtained by clustering the NAS delay states into N groups using the k-means clustering algorithm. The output of the clustering algorithm represents the closest typical state to each of the observations, where the "typical states" are given by the centroids of each of the clusters.

3.1.1 Clustering algorithm

The k-means algorithm partitions n observations into k clusters which minimize the sum of distances within each cluster. The following objective function is minimized:

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

where x_i are the observations and the c_i centroids of each of the clusters. The MATLAB function *kmeans()*, available in the statistics toolbox, was used the solve the clustering problem.

3.1.2 Number of clusters

In this section, we study the effect of the number of clusters on the performance of the clustering algorithm, and the characteristics of the clusters. We then evaluate the value of adding more clusters by looking at the total point-to-centroid Euclidian distance reduction, as well as the physical implications of the clusters.

Figure 3-1 depicts the sum of the distances from each of the clusters' centroid to each of the observations belonging to that cluster for different numbers of clusters (x-axis): This metric is called the intra-cluster distance. Each point contains information associated with one cluster for a certain number of clusters. The color of the dots depicts the number of elements that belong to each of the clusters. We can observe that there is always a significant number of elements belonging to one of the clusters, brown or orange dots. These clusters are associated with low traffic periods. Due to the low delay values during overnight periods a large number of observations fall into this low delay cluster. In Figure 3-2 we can see the sum of the intra-cluster distances for different number of clusters. The total intra-cluster distance gradually decrease as the number of cluster increases. It is worth noting that for more than 4 clusters the slope of the curve reduces significantly, and from 4 to 10 clusters the curve nearly follows a straight line.

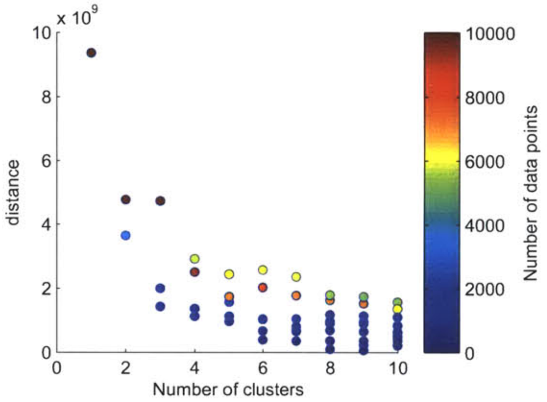


Figure 3-1: Intra-cluster distances vs. number of clusters.

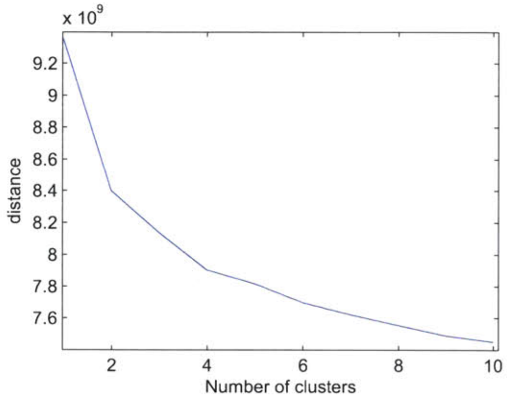


Figure 3-2: Total intra-cluster distance vs. number of clusters.

With the purpose of better understanding the data presented in Figure 3-1 and Figure 3-2 we present Table 3.1. This table contains information about the clusters obtained for different number of clusters, that is: the number of observations belonging to each of the clusters, the average link delay of the clusters' centroid, and a qualitative description of the clusters. As we mentioned previously, there is always a low delay cluster. As the number of clusters increases, the average delay of the low delay cluster decreases; observations with a medium delay level are now associated with other clusters. For 2 clusters, observations where the Chicago (CHI) and New York City (NYC) area are congested belong to the same cluster; however, for 3 clusters, the Chicago-New York cluster splits into two clusters, one with high NYC delays and another one with high CHI delays. For 4 clusters, there is a cluster where CHI is highly congested, another one where NYC is highly congested, the other two clusters show a low and a medium generalized NAS delay, respectively (none of the airports shows up as a congestion center). For more than 4 clusters, the slope of the total intra-cluster distance (see Figure 3-2) does not decrease significantly. However, the results for 5 and 6 clusters are important from our perspective, since ATL appears as a new delay center for 6 clusters. Such a clustering allows us to identify when either NYC, CHI or ATL delays are the main source of delay in the NAS. Between 6 and 10 clusters, the new additional clusters exhibit the same three main sources of delay.

Table 3.1: NAS delay state clustering. Delay definitions: High (90 min), Medium-high (60 min), Medium (20 min), Low (5 min).

Number of clusters	Number of elements	Avg. centroid link delay (min.)	Qualitative Description
2	3,571	30.5	CHI and NYC medium high delay
	13,948	9.7	NAS low delay
3	13,894	9.7	NAS low delay
	1,616	29.6	CHI high delay
	2,009	30.7	NYC high, ATL medium delay
4	9,426	6.6	NAS low delay
	1,219	31.9	CHI high delay
	5,681	17.7	NAS medium delay
	1,193	35.1	NYC high, ATL medium high delay
5	762	38.8	NYC high, ATL medium high
	2,241	23.7	NYC medium high delay
	6,180	13.6	NAS medium delay
	7,099	5.3	NAS low delay
	1,237	31.6	CHI high delay
6	5,915	15.2	NAS medium delay
	1,505	24.4	NAS medium high delay
	1,192	31.2	CHI high delay
	8,029	5.8	NAS low delay
	480	42.2	NYC high, ATL, CHI medium delay
	398	32.9	ALT high delay

In the rest of this chapter and thesis, the analysis of the NAS typical states for the 2007-2008 data is based on clustering the NAS states into six groups. Six clusters are chosen for two reasons: First, the total intra-cluster distance does not decrease much for more than 6 clusters (as seen in Figure 3-2), and second, six appears to be qualitatively reasonable since all the main delay centers (NYC, CHI, ATL) are represented in the centroids of clusters.

3.1.3 Six classes of NAS delay states

In this section we analyze the six most typical delay pictures of the NAS, which are obtained by aggregating the NAS delay states into six groups. Figure 3-3 shows each of the clusters' centroid delay values. The clusters can be characterized as follows:

- **Cluster 1**: Medium delay levels in the entire NAS, with an average link delay of 15.2 minutes, and standard deviation of 4.3 minutes.
- **Cluster 2**: Medium high NYC delays, with an average NAS link delay of 24.4 minutes, and standard deviation of 14.7 minutes.
- **Cluster 3**: High CHI delays, with an average NAS link delay of 31.2 minutes and standard deviation of 23.8 minutes.
- **Cluster 4**: Low delay levels in the entire NAS, with an average link delay of 5.8 minute, and standard deviation of 3.7 minutes.
- **Cluster 5**: High NYC delays and medium ATL and CHI delays, with an average NAS link delay of 42.2 minutes and standard deviation of 31.5 minutes.
- **Cluster 6**: High AL delays, with an average NAS link delay of 32.9 minutes and standard deviation of 25 minutes.

One way to get a sense of how well a cluster' centroid is reflecting observations belonging to that cluster is to look at the location of the observations around the centroid. We have a good cluster if most of the observations are located close to the centroid and the number of observations decreases as we increase the distance from the cluster centroid. However, if the observations are uniformly distributed around the centroid the cluster may need to be split in two clusters, since observations are not gathered around the centroid. Figure 3-4 shows the histogram of the observation-to-centroid distances for each of the clusters. The six figures show that the observations are clusteres around the centroids.

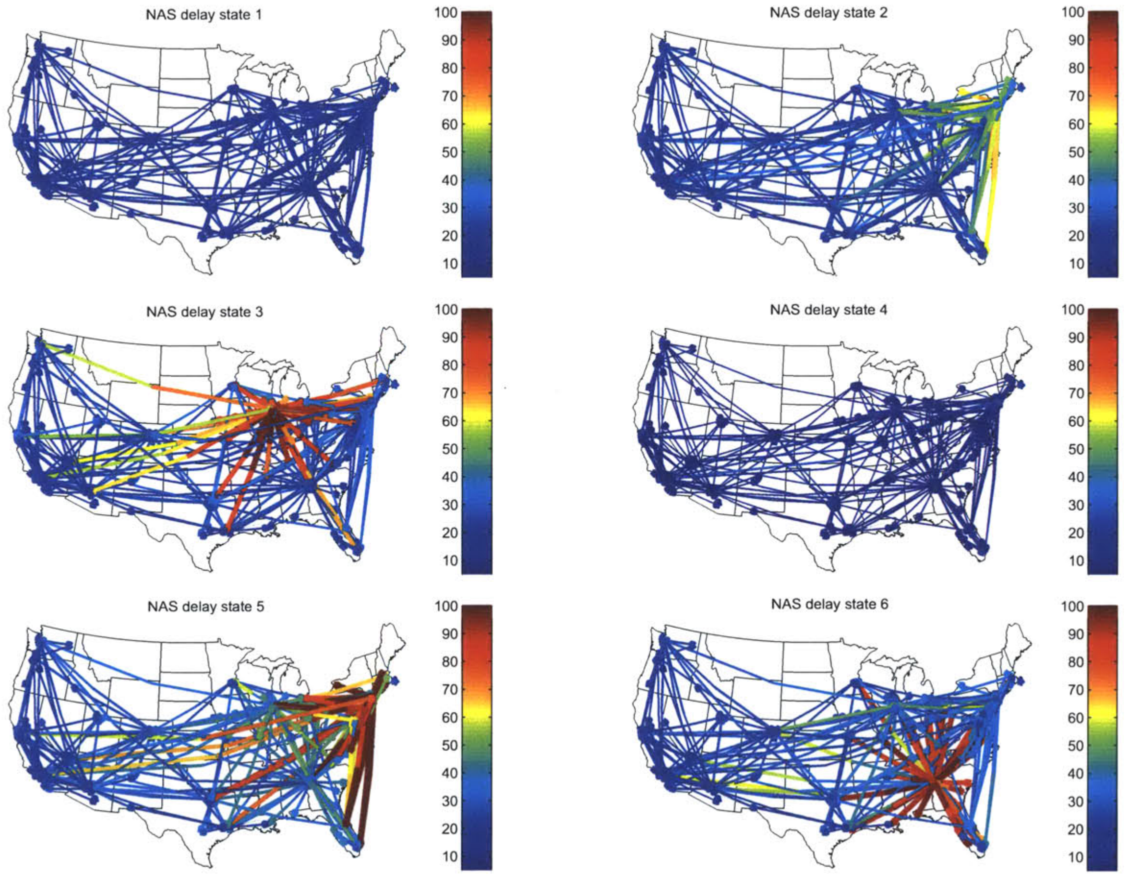


Figure 3-3: Centroids of NAS delay states for six clusters.

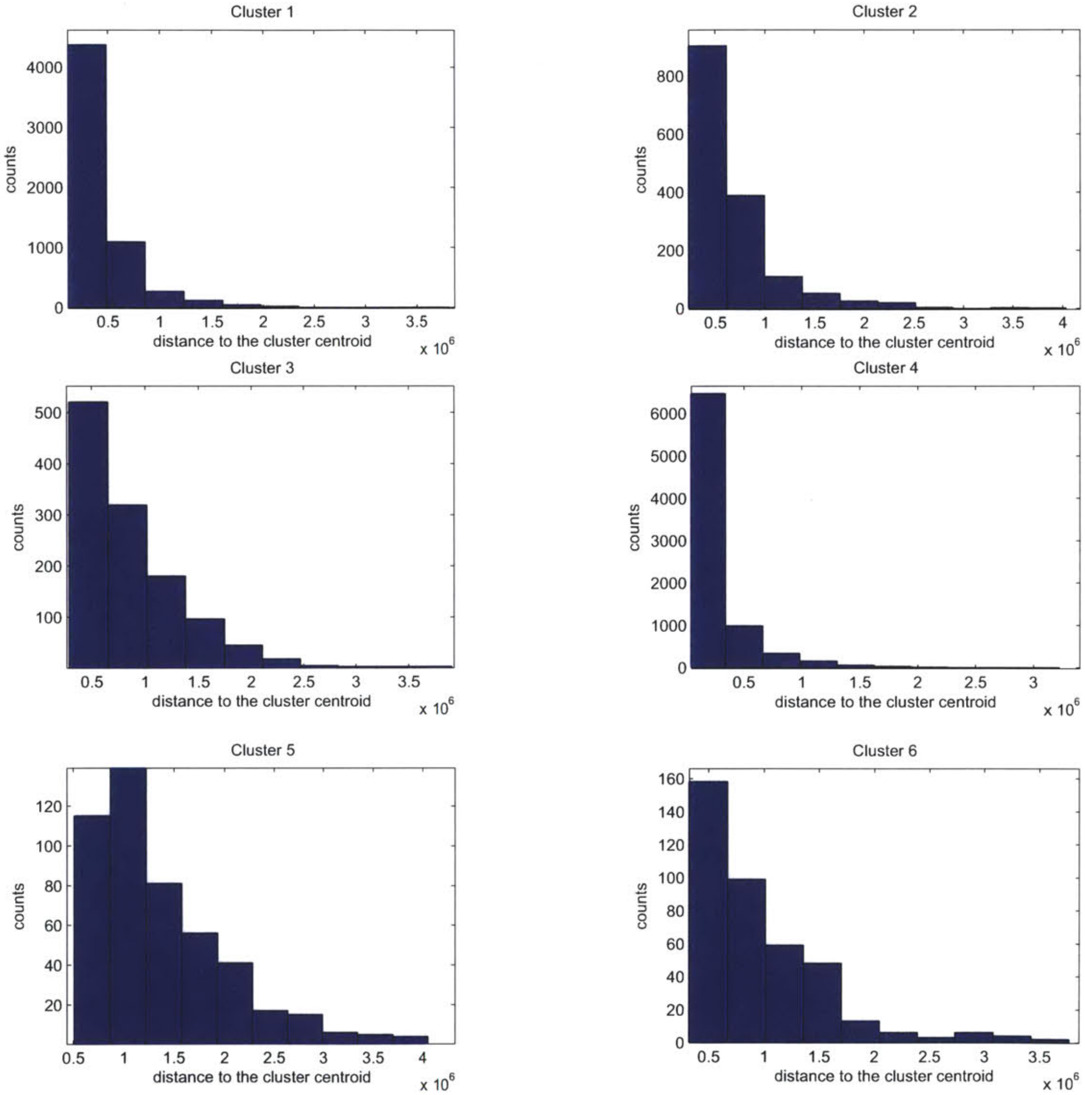


Figure 3-4: Histogram of the observation-to-centroid distances for Clusters 1-6

3.1.4 Analysis of the high-delay clusters

The high delay states introduced previously (clusters 2, 3, 5, and 6) are analyzed in more detail in this section. Our purpose is to determine which airports and links show the highest delays. To do so, we look at the delays associated with the clusters' centroids. It is important to remember that we are analyzing the simplified network with only links that average more than 10 flights a day.

Table 3.2 shows the airports with the highest average departure delay for each of the high delay clusters. Only airports with a significant amount of traffic were considered, as described in Table 2.1. In cluster 2, the NYC airports (JFK, LGA and EWR) are the most congested, along with Philadelphia airport (PHL). In cluster 5, which is the high delay cluster for NYC area airports, we have the same set of airports we had in cluster 2 with delay values that are nearly twice as high. Cluster 3 is the high CHI delay cluster. The difference between ORD delays and the delay of other airports is significant: the average departure delay for arrivals decreases from 88 minutes (ORD) to 41 minutes (EWR). Finally, in cluster 4 ATL has the highest delay levels and there is a significant difference between ATL delays and those at the second most delayed airport, CLT.

These results show that there is a strong correlation among EWR, JFK, LGA and PHI delays (cluster 5). The reason is likely the close spatial location of these four airports, and the resulting correlation in weather.

Next, we look at the clusters centroids' delays at the OD pair level. Table 3.3 shows the ten links with the highest delays for each of the high delay clusters. In cluster 2, NYC congested medium, it is remarkable that nine out of the ten OD pairs contain EWR airport. The remaining OD pair is BOS-JFK. This is consistent with the values we saw in Table 3.2 in which EWR had the highest delay, followed by JFK. It is also worth noting that only two out of the ten OD pairs are departures from the NYC airports. In cluster 5 (NYC congested), the results change. In this case, seven of the OD pairs are either departing or arriving at JFK, and only three of them contain EWR. In cluster 5, we do not see the NYC arrival delay dominance we saw in cluster 2, since only six out of ten OD pairs arrive to NYC airports. It is remarkable that the most delayed OD pairs in cluster 5 have the

Table 3.2: Airports with the highest average link delay in the high delay clusters.

Cluster 2 (NYC congested, medium)			**Cluster 3 (Chicago congested)**		
Airport ID	**Avg link dep. delay(min.)**	**Avg link dep. delay(min.)**	**Airport ID**	**Avg link dep. delay(min.)**	**Avg link dep. delay(min.)**
	Departures	**Arrivals**		**Departures**	**Arrivals**
EWR	55	66	ORD	76	88
JFK	50	52	EWR	39	41
PHL	47	46	DTW	38	35
LGA	45	52	PHL	36	35
BOS	43	44	DFW	35	29
CLT	32	32	LGA	32	34
MCO	31	31	CLT	32	30.5
DTW	31	27	JFK	28	34
Cluster 5 (NYC congested)			**Cluster 6 (ATL congested)**		
Airport ID	**Avg link dep. delay(min.)**	**Avg link dep. delay(min.)**	**Airport ID**	**Avg link dep. delay(min.)**	**Avg link dep. delay(min.)**
	Departures	**Arrivals**		**Departures**	**Arrivals**
EWR	103	104	ATL	87	84
PHL	102	92	CLT	44	41
JFK	98	109	PHL	39	39
LGA	97	97	EWR	38	40
BOS	82	85	DFW	35	30
MCO	62	60	MCO	33	34
CLT	57	62	ORD	33	33
DTW	50	49	DTW	32	29

highest delays among all clusters. With respect to the CHI congested cluster, cluster 3, it is interesting that all the links in the table have ORD as destination. Flights going into ORD have higher delays than flights departing from ORD, as we saw in Table 3.2 (76 min average delay for departures, and 88 min average delay for arrivals). Since it is possible that airlines are aware of ORD's high congestion and tend to schedule tight turnaround times. Finally, in cluster 6 (ATL congested) seven out of the ten OD pairs are arrivals to ATL, the rest are departures from ATL, and the delay level is similar to ORD OD pairs' delay levels. An important point to highlight is that for any of the clusters, the top 10 most delayed links always contain as origin or destination, the NYC airports (EWR, JFK LGA), ATL or ORD.

We also find that ATL and ORD delays are significantly higher that the second-most delayed airport in their respective clusters. We did not see this behavior in the NYC congested clusters. We would like to answer the following question: Do ORD and ATL delays have a lower impact on delay in the rest of the NAS than NYC delays? Scatter plots of 2007-2008 observations are shown in Figure 3-5 to Figure 3-10. In Figure 3-5, we have the

Table 3.3: Links with the highest delay in the high delay clusters

Cluster 2 (NYC congested, medium)			**Cluster 3 (Chicago congested)**		
Origin	**Destination**	**Dep. delay (min.)**	**Origin**	**Destination**	**Dep. delay (min.)**
IAD	EWR	87	STL	ORD	119
CLT	EWR	78	IND	ORD	115
ORD	EWR	78	CLE	ORD	112
BOS	JFK	77	EWR	ORD	111
BOS	EWR	76	SDF	ORD	110
EWR	IAD	75	MSN	ORD	108
DTW	EWR	74	CID	ORD	108
CYYZ	EWR	71	JFK	ORD	107
EWR	CTL	70	MEM	ORD	104
MCO	EWR	69	BNA	ORD	103
Cluster 5 (NYC congested)			**Cluster 6 (ATL congested)**		
Origin	**Destination**	**Dep. delay (min.)**	**Origin**	**Destination**	**Dep. delay (min.)**
BUF	JFK	154	RIC	ATL	111
BOS	JFK	149	DFW	ATL	107
JFK	BOS	143	HOU	ATL	103
MCO	JFK	140	IND	ATL	100
EWR	CLT	132	IAD	ATL	100
EWR	ATL	131	DCA	ATL	100
ATL	EWR	130	ATL	RIC	100
ORD	JFK	130	BWI	ATL	98
JFK	ORD	130	ATL	CLT	98
FLL	JFK	129	ATL	RSW	98

average ORD link delay versus the average link delay on the rest of the network; the color of the dots depicts the average delay at EWR, LGA and JFK. Figure 3-7 shows a similar plot for ATL. We present two 2D histograms for the same x-y information (Figures 3-6 and 3-8), and the NYC scatter plot (dots are not color coded in this plot) and 2D histogram (Figures 3-9 and 3-10, respectively). By comparing the ORD and NYC 2D histograms, we see a higher concentration of data points at the bottom right and upper left side of the ORD figure, and also a lower concentration of points at the upper right side in this same figure. This suggests that ORD delays are less correlated with the network delays than NYC delays. The ORD color coded scatter plot (Figure 3-5) also shows that NYC delays gradually increase as the network delay increases. Finally, the ATL plots look similar to the ORD plots, but with a significantly lower number of high delay data points. There is a significant number of data points at which network delays are high and ATL delays are not, and data at which network delays are low and ATL delays are high.

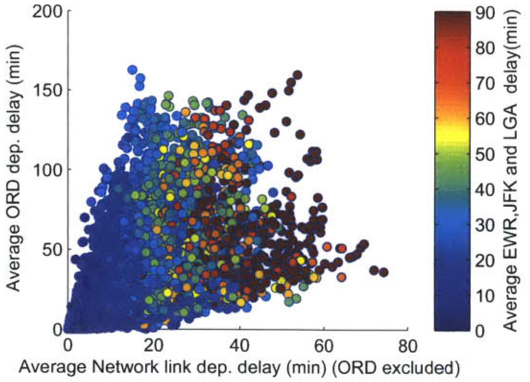


Figure 3-5: Scatter plot of ORD delay vs. rest of the network delay.

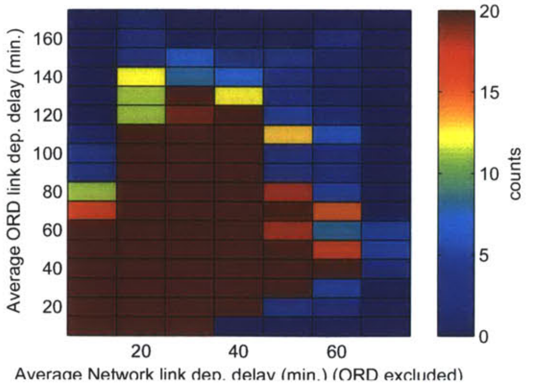


Figure 3-6: 2d histogram of ORD delay vs. rest of the network delay.

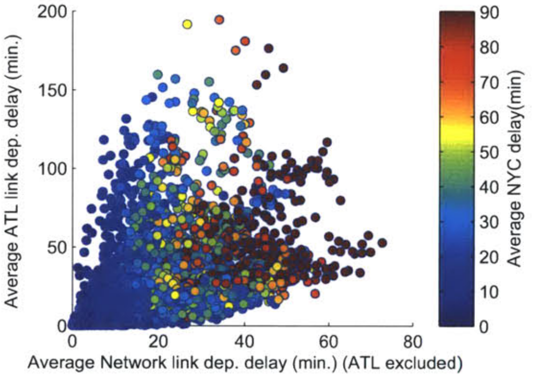


Figure 3-7: Scatter plot of ATL delay vs. rest of the network delay.

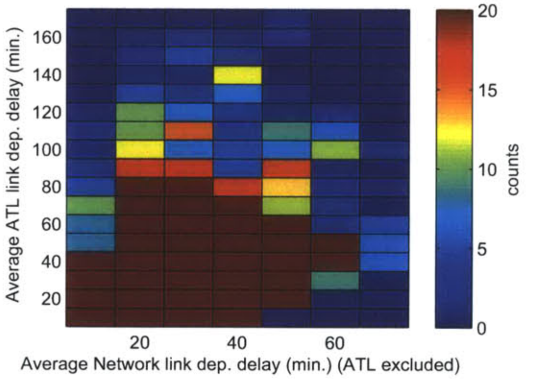


Figure 3-8: 2d histogram of ATL delay vs. rest of the network delay.

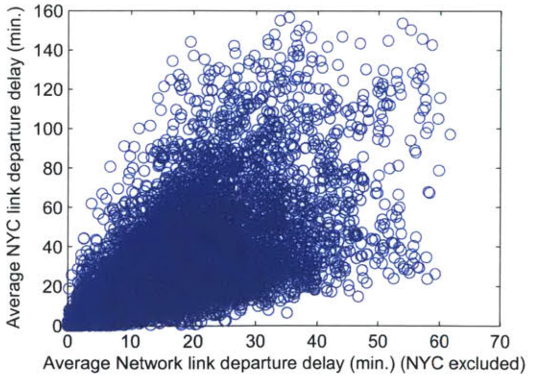


Figure 3-9: Scatter plot of NYC (JFK, EWR, LGA) delay vs. rest of the network delay.

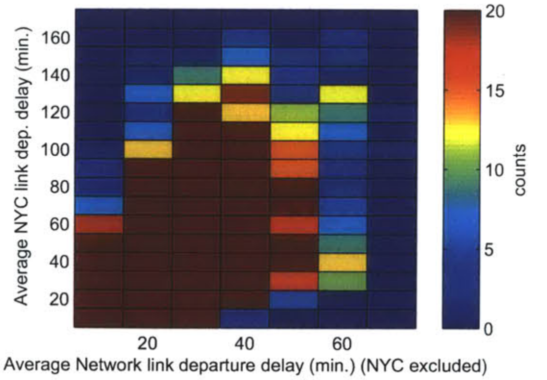


Figure 3-10: 2d histogram of NYC (JFK, EWR, LGA) delay vs. rest of the network delay.

3.1.5 Temporal analysis of NAS delay states

In this section, we study the temporal evolution of global NAS delays. The goal is to get a sense of when, and how often the typical NAS delay state we analyzed earlier occur. To accomplish this, we look at the observations associated with each of the six typical NAS states and the time of day those observations occured.

Figure 3-11 depicts the average NAS delays (one hour window) for January 2007. This figure shows the average link delay for each observation, and the average link delay of the centroid of the cluster that each observation belongs to. The average centroid delay takes one of six different values, one for each cluster: 5.8, 15.2, 24.4, 31.2, 32.9 and 42.2 minutes. Figure 3-11 shows that on some days, the average network link delay goes above the average centroid delay of the 3 highest NAS delay states (ORD, ATL or NYC highly congested); by increasing the number of clusters we could decrease the differences between the red and blue curve.

For the two years of data studied the three highest delay states were reached on 342 out of 730 days; if we also add cluster 2 (NYC medium high delay) we have that those four states were reached on 506 out 730 days (69% of the days).

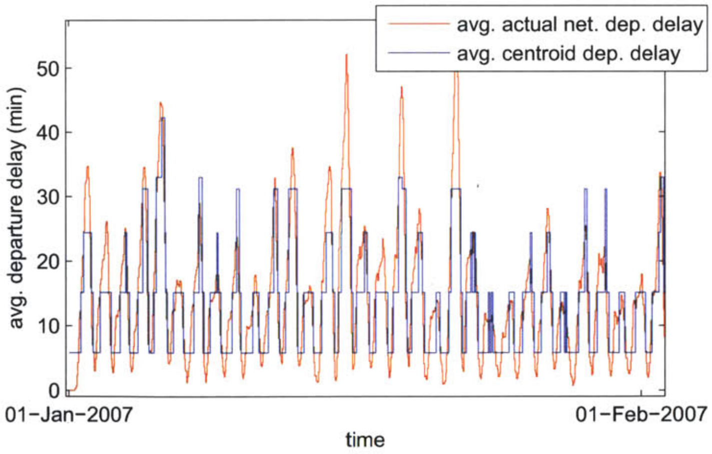


Figure 3-11: NAS average departure delay (January 2007)

Table 3.4 and Figure 3-12 present the analysis of the NAS state occurrences by day and time-of-day, respectively. The data shows that the states with the lowest average delays are the most common; state 4 (NAS low delay) is the most common at night time and state 1

(NAS medium delay) during the day. As we could expect, Figure 3-12 shows that the slope of state 4 curve is higher at the end of the day (rising edge) than at the beginning of the day; at night time there is a point at which the number of departures drops and delays also drop, but when the day begins delays increase progressively. State 2 (NYC congested medium) and state 3 (ORD congested) counts start increasing at the same time of day, around 11am eastern time. However, states 5 and 6 (NYC and ATL high delay clusters) do not start occurring until 5pm. One factor leading to the late start of the NYC high delay state is that state 2 (NYC congested medium) is a transition state; delays in NYC area start increasing and the system reaches state 2, they continue increasing until state 5 is reached. In the Chicago case, when delays in ORD are significant the system transition directly to state 3 (ORD congested), there is no intermediate ORD state. Finally, at the end of the day, all the high delay states counts decrease past midnight.

Table 3.4: NAS states statistics.

NAS States	Percentage of days each state occurs	Avg. active time, if active (hours)
State 1	100%	8.2
State 2	50%	4.1
State 3	29%	5.8
State 4	100%	11
State 5	16%	4.1
State 6	12%	4.6

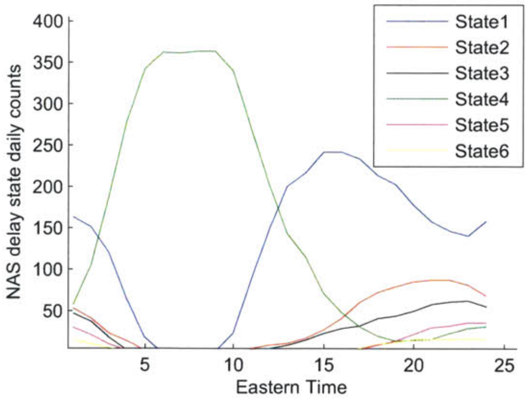


Figure 3-12: State occurrences by time of day (2007-2008 data).

Figure 3-13 shows the monthly occurrences of each of the NAS states. In this figure we note that Chicago high delay state (state 3) takes place more often in the winter months, and that NYC and ATL high delay states (state 5, and 6) are more frequent in the summer months. Figure 3-13 also suggests that September, October, and November are the months in which less high delay states take place, and in which the low delay state (state 4) is more frequent.

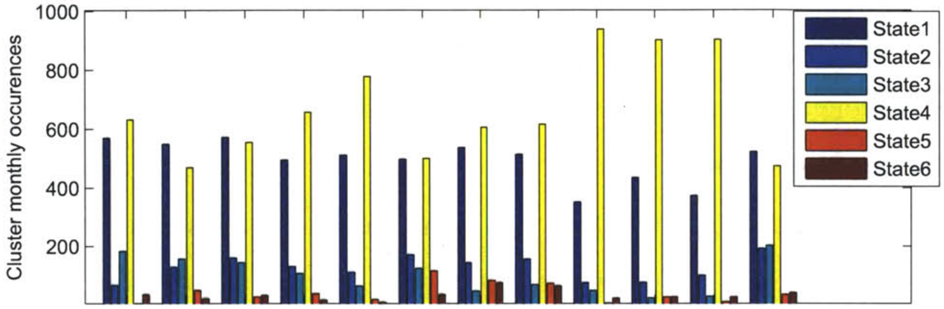


Figure 3-13: Monthly occurrences of NAS state.

3.2 NAS type-of-day

In addition to clustering the NAS delay state at time t, we cluster entire days of time series data. The idea is to identify a set of typical NAS type of days according to the daily delay of all the links in the simplified network. Each of the data points we cluster has $584\times24=14{,}016$ variables (Number of OD pairs × 24 hours, we have one observation per hour due to the 1-hour step size of the moving median filter).

Figure 3-15 shows the total within cluster distance for different numbers of type-of-day clusters, and Figure 3-14 the within-cluster distance for each of the clusters. These results were obtained using the k-means algorithm.

We followed the same methodology presented in the previous section to choose the number of NAS state clusters (based on distance reduction and qualitative description of the centroids), and we chose six clusters again. Instead of a video to visualize the clusters' centroids, we present Table 3.5 which describes the main source of delay at the highest delay point of the day for different numbers of type-of-day clusters. The average daily delay and the number of observations belonging to each cluster are also included. The

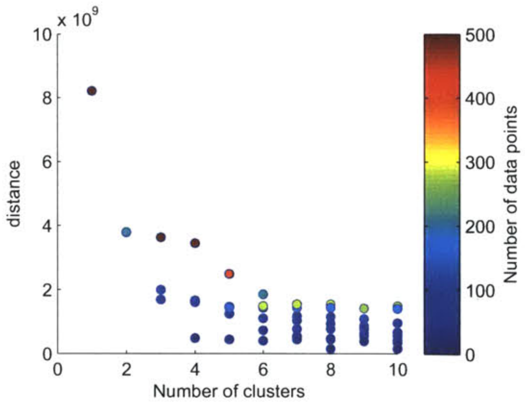


Figure 3-14: Intra-cluster distances vs. number of clusters.

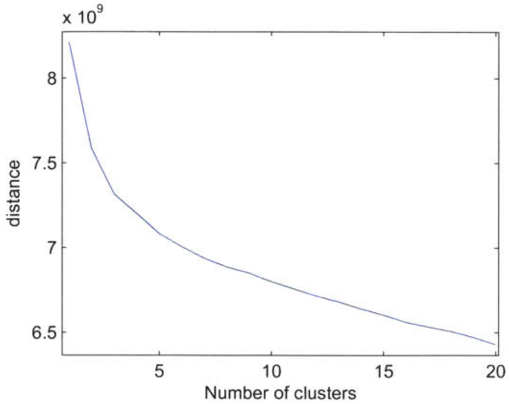


Figure 3-15: Total intra-cluster distance vs. number of clusters.

main delay sources are the same we saw in the NAS state clusters: ORD, ATL, and NYC. Five clusters would allow us to identify the main delay sources in the NAS; however, we choose 6 clusters because the distance reduction from 5 to 6 clusters is significant.

3.2.1 Temporal analysis of NAS type-of-day

In this section, we study the monthly frequency of each of the six types of day identified in the previous section. Similar to the analysis of the NAS states, our goal is to identify any significant seasonal patterns.

Figures 3-16 shows the monthly occurrences of the different types of days. We see that Day 1 (high NYC delays, and significant ORD and ATL delays) is more common in the summer months, while Day 6 (high NYC delays, but not high ORD or ATL delays) is seen year-round, with higher frequency in the summer months. We also see that the Chicago high delay day (Day 2) is more frequent in the winter, while the Atlanta high delay day (Day 4) is more frequent in the summer.

It is interesting that the NAS state monthly occurrences, and the type of day monthly occurrences show very similar results, likely because the same delay sources define them. For example, if we have an ATL high delay day, most of the high delay NAS states of that day will be ATL high delay states. It is unlikely that we will have many other high delay states in a day classified as on ATL high delay type-of-day; we could see some other high delay states, for example, the ORD high delay state, in the transitions from high to low

Table 3.5: NAS type of day clustering. Delay definitions: High (90 min), Medium-high (60 min), Medium (20 min), Low (5 min).

Number of clusters	Number of elements	Avg. centroid link delay (min.)	Qualitative Description
2	521	11.9	NYC medium delay
	208	23.1	CHI and NYC medium high delay
3	107	22.3	CHI high, NYC medium high delay
	508	11.8	NYC medium delay
	114	22.9	NYC high, ATL, ORD medium high delay
4	104	22.1	CHI high, NYC medium high delay
	104	21.6	NYC high, ATL, ORD medium delay
	494	11.7	NYC medium delay
	27	25.4	ATL high, NYC, ORD medium high delay
5	61	25	NYC high, ATL, ORD medium high delay
	31	21	ATL high, NYC, ORD medium high delay
	97	22	CHI high, NYC medium high delay
	397	11	NAS low delay
	141	17	NYC high, ATL, ORD medium delay
6	31	29	NYC high+, ATL, ORD high delay
	94	22	CHI high, NYC medium high delay
	207	15	NYC, ORD medium delay
	29	21	ATL high, NYC, ORD medium high delay
	282	9	NAS low delay
	86	19	NYC high, ATL, ORD medium delay

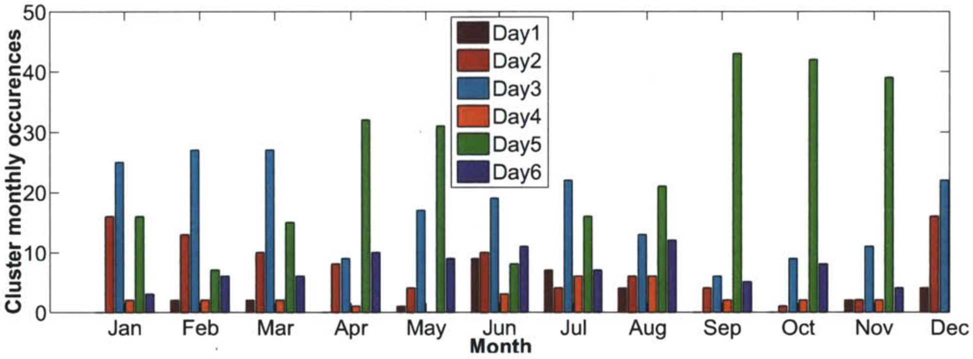


Figure 3-16: Monthly occurrences of NAS type-of-day.

delay states, or vice versa.

Chapter 4

Analysis of Explanatory Variables

The goal of this chapter is to identify those variables that can play an important role in predicting the level of delay of a certain link in the air traffic network. We will predict the departure or arrival delay state of a link at time $t+T$, using temporal variables (e.g. time-of-day, day-of-week or season), the local delay state variables value at time t (which depict the most relevant airports' and links' delay states) and the high level delay state variables value at time t (these are the categorical variables obtained in Chapter 3 through clustering). The final goal will be to estimate a measure of the departure or arrival delay state of a link a few hours into the future.

The analysis of the different variables presented in this chapter will focus one specific link. The selected link for this analysis is JFK-ORD, which had the highest delay values in the NAS in the 2007-2008 period. It also connects two main delay centers, New York City and Chicago.

The rest of the thesis focuses on predicting departure delays. A similar analysis can be done for arrival delays.

4.1 Description of explanatory variables

In this section, we evaluate the effect of changes in different categorical and continuous variables on the JFK-ORD departure delay state. Our goal is to determine whether or not to include the explanatory variables proposed below in our delay prediction models. We will do so by analyzing the effect of changes on those variables in the departure delay state that we are interested in predicting. The analyzed variables can be divided into two groups:

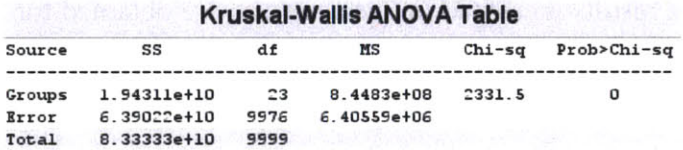
1. Temporal explanatory variables: Time-of-day, day-of-week, season.
2. Network delay state explanatory variables:
   - Local delay state variables: Influential airports delay state, influential OD pairs delay state.
   - High level delay state variables: NAS delay state, NAS type-of-day, previous day's type-of-day.

In the categorical variables analysis presented below, we the ANOVA test, and the multiple comparisons test as the tools to evaluate the dependence of the departure delay with the different categories of the proposed set of variables [12]. The ANOVA test provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes the t-test to more than two groups. Multiple comparisons allows us to evaluate differences among the different categories, which the ANOVA test does not show since it looks at all the groups as a whole (as long as the means of two categories different, the ANOVA test will state that the means of the different categories are different). We used a non-parametric ANOVA test (Kruskal-Wallis test) due to the skewness of the delay distribution, which does not satisfy the normality assumption required by the parametric ANOVA test. The delay distribution is skewed to the left because we have more low delay points than high delay points in our data set. As a consequence of the non-parametric ANOVA test, all the multiple comparisons tests presented here are based on ranks, and they were calculated using the Tukey-Kramer criterion, and a 95% confidence interval. 5,000 data points where randomly sampled with replacement from the two years of data (17,519 observations) to performing the statistical analysis presented in the following sections. A different approach was followed in the identification and analysis of the continuous variables, where we used a Random Forest (RF) based methodology to identify the relevant delay variables. More details are presented in Section 4.1.6.

The significance of the explanatory variables presented in this research depends on the specific output of the prediction model we consider (regression vs classification). In the analysis presented in the following chapters, we evaluate two types of prediction models: classification models, where the output is a binary prediction that indicates whether the delay level is higher or lower than a predefined threshold, and regression models, where the continuous output directly estimates the delay along a link of interest. We assume that if an explanatory variable has an effect on the continuous delay output, it will also have an effect on the binary variable; for this reason, we only consider the continuous delay output in this chapter. The results presented in this chapter were obtained for a 2-hour prediction window, but some interesting results for a 4-hour window are also included.

4.1.1 Time-of-day explanatory variable

In this section, we study how the JFK-ORD departure delay state varies over the course of the day. The time-of-day is a continuous variable; however, we are going to treat it as a categorical variable by defining 24 time periods, one for each hour of the day. The Kruskal-Wallis test table is presented in Figure 4-1, and shows a 0 p-value. The p-value gives the probability of obtaining a more extreme value of the test statistic, assuming that the null hypothesis is true. The null hypothesis is rejected when the p-value is smaller than the significance level α, which is often 0.05 or 0.01. A low p-value means that the null hypothesis is very unlikely to be true for our data set. The ANOVA test null-hypothesis is that all categories have equal means. Consequently, a zero p-value means that the null-hypothesis is rejected, an there is a difference between the mean JFK-ORD departure delay for the different time-of-day categories. Figure 4-2 shows the multiple comparisons intervals. We see that as the day progresses delays increase, and around 3am delays start decreasing. It is important to remember that our data depicts the departure delay of flights *d*eparting at time *t*; consequently, flights departing at 3am from New York are very likely to be highly delayed. At some points of the day, there is no a significant difference in consecutive hours' delays (e.g. 15h to 16h). We could aggregate some of these categories, resulting in the same information but fewer categories, which would decrease our prediction models' complexity. However, we keep 24 categories in order to keep the maximum level of detail in the time-of-day information. For our prediction models, it is going to be very hard to predict delays when they start increasing, or when they start decreasing. The time-of-day variable plays a very important role here, by determining when delays typically start increasing or decreasing. Secondly, we want to develop a prediction model with a common structure that can be used in different links. The time-of-day variable effect on the departure delay can change from link to link. We would need to analyze each of the links in the network individually if we aggregate the time-of-day variable categories.



```
                      Kruskal-Wallis ANOVA Table
Source        SS         df         MS         Chi-sq   Prob>Chi-sq
---------------------------------------------------------------------
Groups   1.94311e+10     23     8.4483e+08    2331.5          0
Error    6.39022e+10   9976    6.40559e+06
Total    8.33333e+10   9999
```

Figure 4-1: Time-of-day ANOVA table for JFK-ORD departure delays.

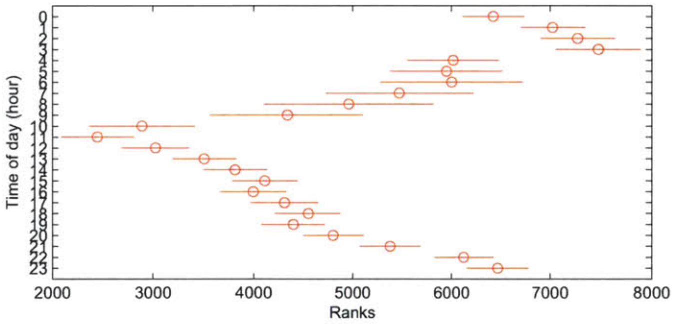
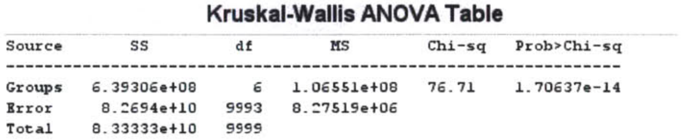


Figure 4-2: Time-of-day multiple comparisons test for JFK-ORD departure delays.

Finally, if we take a look at the width of the confidence intervals in Figure 4-2 we see that they are wider at the end of the day, when delays decrease. This happens because, at that time, there is typically no congestion, but sometimes flights are still propagating large delays from the previous day.

4.1.2 Day-of-week explanatory variable

We start by evaluating 7 categories, one for each day of the week. The question to answer is: Do JFK-ORD departure delays differ by day of the week? In Figure 4-3, we show the K-W ANOVA test p-value, which is very close to zero, and Figure 4-4 (category 1 is Sunday, category 2 Monday, and so on) also shows that delays change for different days of the week, since a good number of the multiple comparisons intervals do not overlap. Friday (category 6) has the highest delay, and Tuesday, Wednesday and Saturday the lowest delay (categories 3, 4 and 7).



Kruskal-Wallis ANOVA Table

```
Source       SS          df        MS          Chi-sq   Prob>Chi-sq
-------------------------------------------------------------------
Groups   6.39306e+08      6   1.06551e+08   76.71    1.70637e-14
Error     8.2694e+10   9993   8.27519e+06
Total    8.33333e+10   9999
```

Figure 4-3: Day-of-week ANOVA table for JFK-ORD departure delays (7 categories).

We can increase the significance of the different values of the day-of-week variable by aggregating the seven days of the week in 3 groups: the low delay category (Tuesday, Wednesday and Saturday), the medium delay category (Sunday, Monday, and Thursday), and the high delay category (Friday). Figure 4-5 shows the multiple comparisons plot for these three groups, which none of the groups overlap. The problem here is that the day of the week affects different OD pairs differently, and we want to define a set of categories that make sense for all the OD pairs, and do not lead to information loss. Figure 4-6 shows the system-wide multiple comparisons plot (all OD pairs' departure delay treated as one single OD pair) for 7 categories of day-of-week variable. We see that the aggregation we proposed for the JFK-ORD link does not make much sense here. For these reason, we choose seven categories for our day-of-week variable.

4.1.3 Season explanatory variable

For the last of the temporal variables, the season, we perform a similar analysis. A low p-value (see Figure 4-7), and multiple comparisons intervals that do not overlap (see Figure 4-8) show that JFK-ORD departure delays change for different months of the year. Figure 4-9 shows the system-wide (all OD pairs treated as a single OD pair) departure delay dependence with the month of the year. Both the JFK-ORD multiple comparisons plot and the system-wide plot depict a very similar delay dependence with the month of the

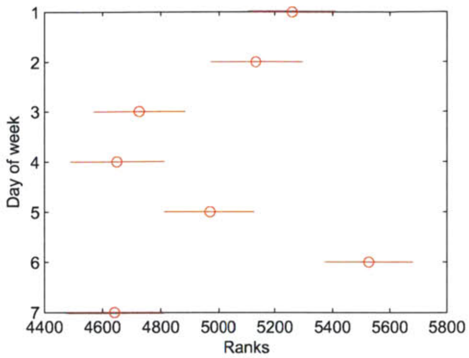


Figure 4-4: Day-of-week multiple comparisons test for JFK-ORD departure delays (7 categories).

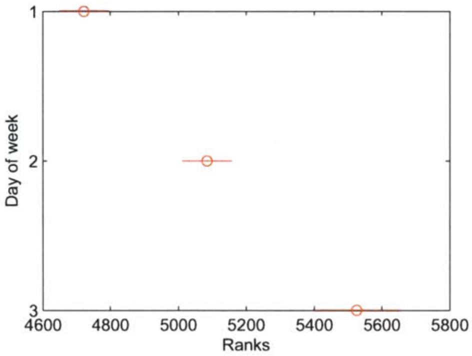


Figure 4-5: Day-of-week multiple comparisons test for JFK-ORD departure delays (3 categories).

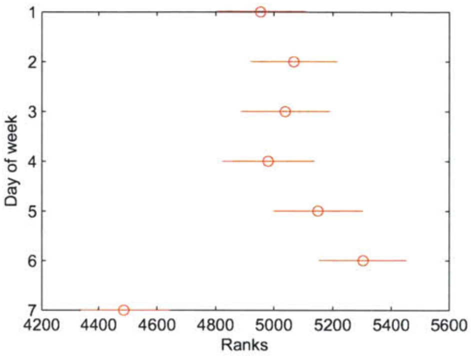
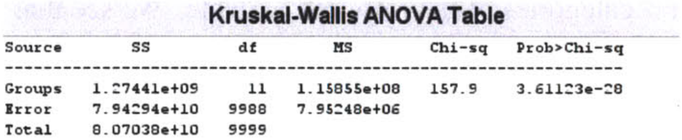


Figure 4-6: Network aggregated day-of-week multiple comparisons test for JFK-ORD departure delays (7 categories).



Kruskal-Wallis ANOVA Table

```
Source        SS          df        MS          Chi-sq    Prob>Chi-sq
-----------------------------------------------------------------------
Groups    1.27441e+09     11    1.15855e+08     157.9     3.61123e-28
Error     7.94294e+10   9988    7.95248e+06
Total     8.07038e+10   9999
```

Figure 4-7: Season ANOVA table for JFK-ORD departure delays (12 categories).

year. It makes sense to aggregate the month of the year into three categories. September, October, and November show up as the low delay months (recall that this is in line with the results of the analysis of the type-of-day and NAS state monthly occurrences, where we identified these three months as the low delay months). January to May can be aggregated as the medium delay months, and December, and the summer months (June, July, and August) can be aggregated as the high delay months. Convective weather and high

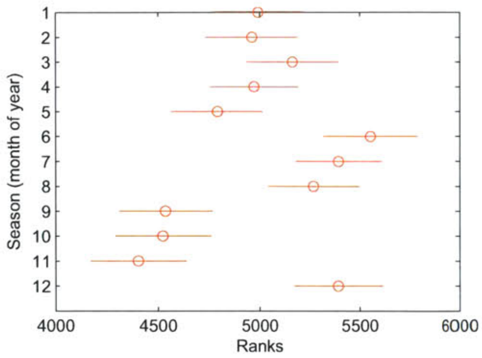


Figure 4-8: Season multiple comparisons test for JFK-ORD departure delays (12 categories).

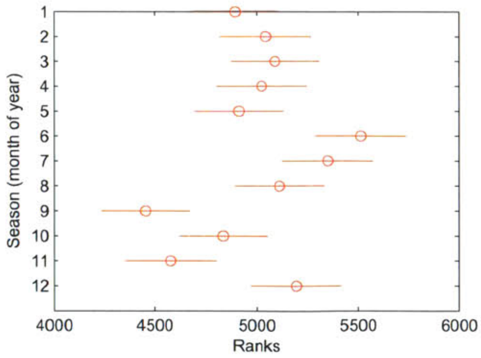


Figure 4-9: Network aggregated season multiple comparisons test for JFK-ORD departure delays (12 categories).

demand levels are the main causes for the high delay months. Finally, Figure 4-10 shows the multiple comparisons plot for the three defined categories for the JFK-ORD OD pair. The three categories lead to significantly different levels of delay.

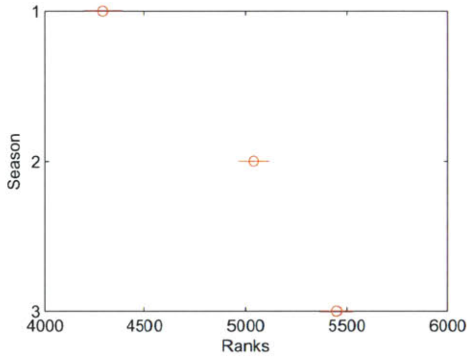
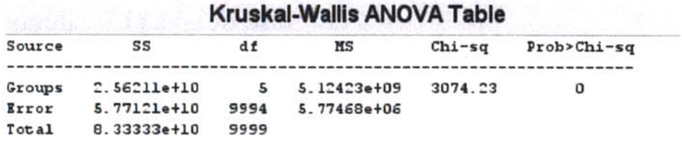


Figure 4-10: Season multiple comparisons test for JFK-ORD departure delays (3 categories).

4.1.4 NAS delay state explananatory variable

In the previous chapter, we identified the most typical delay NAS delay states using the k-means clustering algorithm. The categorical variable presented in this section depicts the NAS typical state that the system is closest to at time t. We evaluate the dependence of the future delay of a certain OD pair on the global delay state of the NAS.

The analysis presented here was performed for 6 clusters. In Figure 4-11, we see that the ANOVA test p-value is 0, omplying that the means of the JFK-ORD departure delay for different values of the NAS state categorical variable are not equal, and in Figure 4-12 we see the associated multiple comparisons intervals. It is reasonable that State 4 leads to the lowest delay interval, since it is the low NAS delay state. The next highest JFK-ORD delays are for State 1, which is the medium NAS delay state. The ATL high delay state (6) comes next: The JFK-ORD delay levels are not too high for this state. The next state is the NYC medium-high delay one (State 2), and finally we have the Chicago and NYC high delay states (3 and 5).



Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.56211e+10	5	5.12423e+09	3074.23	0
Error	5.77121e+10	9994	5.77468e+06		
Total	8.33333e+10	9999			

Figure 4-11: NAS state ANOVA table for JFK-ORD departure delays.

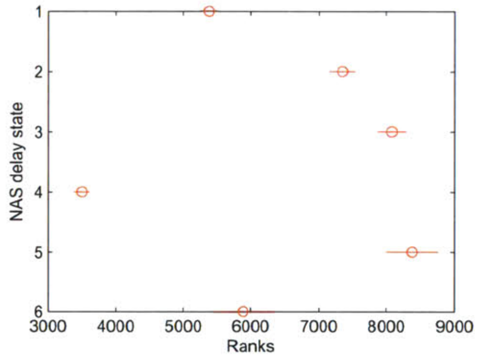


Figure 4-12: NAS State multiple comparisons test for JFK-ORD departure delays.

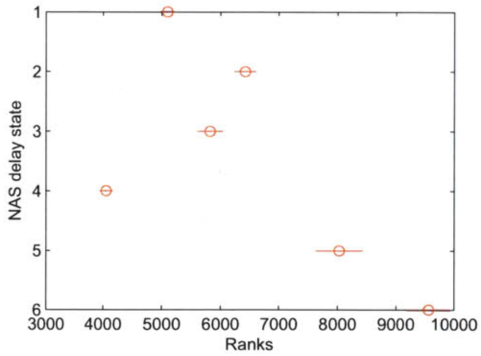


Figure 4-13: NAS State multiple comparisons test for ATL-MCO departure delays.

While some of the NAS delay state categories could be merged in this case, we do not want to make model simplifications that could worsen the model performance on other OD pairs. For example, Figure 4-13 shows the multiple comparisons intervals for the ATL-MCO link. We note that the ATL high delay state is now the highest delay state. We also note that the two low delay states continue being the same. It is also important to highlight that the NYC medium delay state, which has an average network link delay of 24.4 minutes (State 2), leads to higher delays in the ATL-MCO link than the Chicago high delay state (State 3), which has an average network link delay of 34.2 minutes. in other words, NYC area congestion seems to be more correlated with ATL-MCO delays than Chicago congestion. This fact is consistent with our solution in section 3.1.4: the network delay is more correlated with NYC delays than Chicago delays.

Finally, we see how a change in the prediction horizon from 2-hour to 4-hour affects the dependence of the JFK-ORD departure delay on the NAS state categorical variable. Figure 4-14 depicts the JFK-ORD departure delay multiple comparisons plot for a 4 hour time horizon. Comparing Figure 4-14 and Figure 4-12, we see that the overlap and interval width of the three high delay states increases with the prediction horizon. The increase in the interval width indicates more uncertainty in the JFK-ORD delay; it is harder to predict delays 4 hours into the future than 2 hours into the future. There is also a significant separation between the high delay states and State 1 (NAS medium delay state). These results show that on increasing the time horizon, the NAS state categorical variable loses significance, leading to a decrease in predictive power. In Chapter 5, we evaluate the effect of changes in the prediction horizon on the performance of the prediction model (with all explanatory variables included).

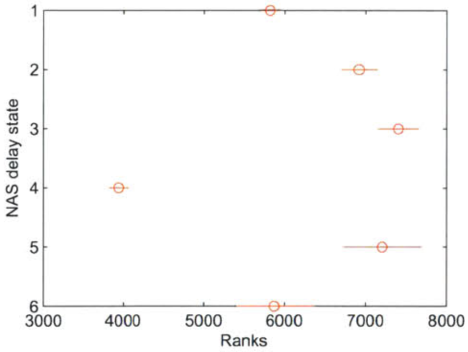
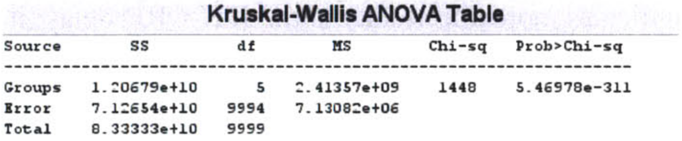


Figure 4-14: NAS state multiple comparisons test for JFK-ORD departure delays with a 4h prediction horizon.

4.1.5 NAS delay type-of-day explanatory variable

We hypothesized that the type of NAS day (as identified in Section 3.2)would have an impact on the future delay level of any given OD pair. For six clusters, we have the results shown in Figure 4-15, and Figure 4-16. We have a very low p-value and differences among the JFK-ORD departure delays for the different categories of the explanatory variable. The NAS low delay day (day type 5) has the lowest delay, followed by the medium delay day (day type 3), then we have the ATL high delay day, which does not have a big impact on the JFK-ORD departure delay, and finally the two type of days in which Chicago or NYC delays are very high, leading to the highest JFK-ORD departure delays in the multiple comparisons plot. These two types-of-day could be merged because the almost perfectly overlap. However, for other links, for example STL-ORD (see Figure 4-17), both types-of-day are important since they lead to different delay levels.



```
                    Kruskal-Wallis ANOVA Table
Source        SS          df        MS          Chi-sq   Prob>Chi-sq
--------------------------------------------------------------------
Groups   1.20679e+10       5   2.41357e+09     1448    5.46978e-311
Error    7.12654e+10    9994   7.13082e+06
Total    8.33333e+10    9999
```

Figure 4-15: Type-of-day ANOVA table for JFK-ORD departure delays.

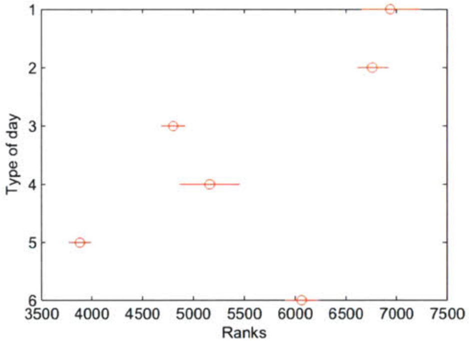


Figure 4-16: Type-of-day multiple comparisons test for JFK-ORD departure delays.

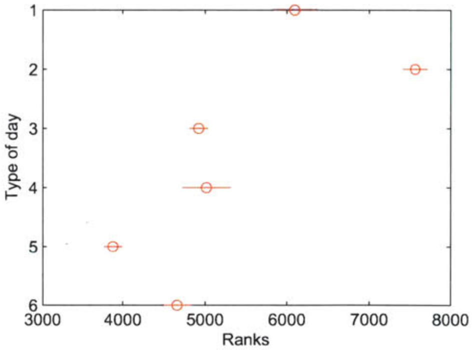


Figure 4-17: Type-of-day multiple comparisons test for STL-ORD departure delays.

We note that one needs the entire days' delay information to determine the type of a given day. In practice, if we make a delay prediction at 2 pm, we only have the delay information from the beginning of the day to 2 pm. Although the type of day should be estimated with the information available at the time the prediction is made, we assume that we know the type of day with certainty before the day is over. While evaluating the prediction capabilities of the type-of-day variable, we do not include the errors in estimating it.

NAS delays for the previous day are known with certainty, and can help predict delays later in the day. The NAS does not immediately recover from high delay situations, such as, a day with strong convective weather or a large number of canceled flights. Passengers will be accommodated in flights over the next few days, leading to higher traffic levels and subsequent delays. Scheduled aircraft routings are also affected by canceled flights, causing additional delays. Figure 4-18 shows the previous days' type-of-day multiple comparisons plot for the JFK-ORD departure delay. As we expected, if the previous day suffers high delays (days type 1 and 2) delays tend to be higher on the next day than if the previous day's delays were low (day type 5). Both variables, the NAS delay type-of-day and previous day type-of-day, are included and evaluated in the prediction models presented in the following chapters.

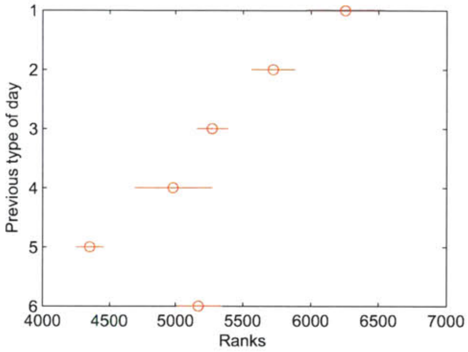


Figure 4-18: Previous day type-of-day multiple comparisons test for JFK-ORD departure delays.

4.1.6 Influential airports explanatory variables

The influential airports for a given delay prediction problem are those airports whose arrival or departure delay states play an important role in predicting the delay of the OD pair of interest. In this section, we want to identify the airports in the network that can help predict the ORD-JFK departure delay. It seem evident that the ORD departure delay, and the JFK arrival delay will play an important role here; however, could any other airports delay variable help us make a better prediction?

In the reduced network, we have 103 different airports. These are airports at which at least one of the OD pairs operates 10 flights a day on average. We extent the set of possible influential airports to the 400 airports with the most traffic; in order to sure that we do not omit any significant airports. We analyze departure and arrival delays separately, resulting in 800 variables in our problem. For the specific OD pair that we analyze in this chapter (JFK-ORD), we identify which of those 800 variables play the most important role in predicting the departure delay 2 hours into the future. The variable selection method presented in this thesis is based on the variables' importance level obtained from running a regression Random Forest (RF) algorithm. The RF provides a good measure of the importance of each of the variables in the model, and works well if the the number the variables is large compare to the number of samples. More details can be found in [13]

With the purpose of increasing the robustness of the variables importance, and to avoid very high computational times, we follow the following methodology to identify the set of airports of interest:

1. Sampling, with replacement, of 3000 training data points from the 2007-2008 data set, and fitting of a RF with 15 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 10 different training data sets with 3000 samples (w.r.) each, and fitting of a RF with 100 trees to each of those training data sets. The final variable importance values will be the average of the values obtained from the 10 RFs.

The training sets were not randomly sampled from the 2007-2008 data; instead we used over-sampling. Over-sampling is the "deliberate selection of individuals of a rare type in order to obtain reasonably precise estimates of the properties of this type. In a population which includes such a rare type, a random sample of the entire population might result in very few (or none) of these individuals being selected" [14]. Over-sampling allows us to have a balanced data set, and to therefore avoid having more low delay data points in our training sets. This is especially important in a classification problem: If we want to classify future delays as high (e.g., over 60 min) or low (under 60 min), we want half of the points in our training and test sets to present delays of over (or under) 60 min. We will use over-sampling in the rest of the thesis in both training and test sets.

The results of the proposed algorithm for the JFK-ORD departure delay link are shown in Table 4.1. This table shows the 10 airports with the highest importance level. As we were expecting the JFK departure delay and ORD arrival delay variables have high importance levels, 96.9 and 85.3 respectively. However, the DCA arrival variable has the highest importance value. There are two possible reasons for DCA departure delay variable high importance in Table 4.1. The first reason is the airport location: DCA is located slightly south of the line that connects JFK and ORD airports. Weather events that take place in between JFK and ORD, which will have a big impact on the JFK-ORD route departure delays, will also heavily affect DCA airport delays. On the other hand, we have the route structure of the airports as another important element. DCA is a more short haul, East Coast flight oriented airport than JFK or ORD. This fact has an important effect in the selection of the most relevant delay variables, since DCA average delay depicts the local delay state well, and consequently the JFK-ORD delay state. DCA does not have flights connecting to the West Coast, or international flights. However, if we look at JFK routes we see flights going to LAS, LAX, SFO, and even to EGLL (London Heathrow). This argument also explains why IAD, which is located next to DCA, is not in the top 10 airports. IAD is, as JFK and ORD, a more long haul and international-oriented airport than DCA.

Table 4.1: Influential airports for JFK-ORD departure delay prediction.

Airport	Delay Type	Variable Importance
DCA	Departure	100
JFK	Departure	96.9
ORD	Arrival	85.3
ORD	Departure	82.8
LGA	Departure	58.9
BOS	Departure	58.9
PHL	Departure	58.2
EWR	Departure	57.7
JFK	Departure	56.3
DCA	Arrival	46.1

Figure 4-19 shows a geographical representation of the JFK-ORD variables importance.

All the variables with at least an importance level of 15 are included, this leads to 24 variables. The outer circles depict the departure delay variables' importance, and the inner circles the arrival delay variables' importance. This figure shows that the airports location is an important factor driving the airport delay variables importance level; most influential airports are located either in the NYC or Chicago areas Figure 4-20 shows the same type of plot, but for the LGA-FLL route. We see that in this case the LGA area delays are dominant, and none of the airports delay variables in the FLL area are relevant. LGA-FLL departure delays are driven by the delay situation in NYC area.

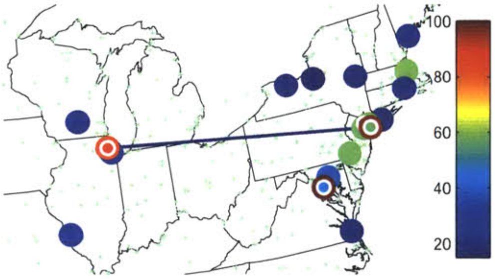


Figure 4-19: JFK-ORD influential airports importance geographical representation. The outer circles depict the departure delay variables importance, and the inner circles the arrival delay variables importance.

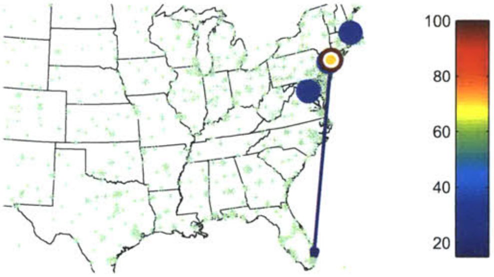


Figure 4-20: LGA-FLL influential airports importance geographical representation. The outer circles depict the departure delay variables importance, and the inner circles the arrival delay variables importance.

4.1.7 Influential OD pairs explanatory variables

In this section, our goal is to identify the OD pairs whose arrival/departure delays can have an important role in a delay prediction model. We use the same methodology presented in the previous section, but instead of airport delay variables we use OD pairs delay variables. We include all the links in the reduced network in our analysis. This leads to 1,064 variables, half of them are arrival delay variables and the other half departure variables.

After running the RF algorithm for the JFK-ORD link with all the links in the reduced network as explanatory variables, we identified the 10 most important links presented in Table4.2. The first thing to notice is that most of the selected variables are departure delay variables. Most flight delays are absorbed on the ground, for this reason departure and arrival delays do not differ much in general. Consequently, the arrival delay of a certain OD pair at time t is close to the departure delay of that OD pair at time t-(flight time). This is why departure delays are typically more valuable information than the arrival delays. However, in some circumstances arrival delays may have a high predictive power; for example, due to connectivity issues, an aircraft that just arrived could be the one departing in a few hours. The need to add more past delay information can also make arrival delays important. In Table4.2, the LGA-ORD arrival delay variable adds more information to the LGA-ORD departure delay variable, which is the third most important variable with a 65.3 importance level. With respect to BUF-JFK arrival delay, over 2007-2008, 302 aircraft flying from JFK to ORD flew into JFK from BUF, possibly explaining the BUF-JFK arrival delay importance (see JFK-ORD aircraft rotation details in Table4.3). The BUF delays are also a good indicator of the weather affecting flights in the area, or airspace congestion issues. Going back to Table4.2, the three most important variables are all associated with NYC airports departure delay of flights going to ORD. The next two variables depict departure delays of flights going from ORD to NYC (JFK and LGA). PHL and BOS departure delays to ORD are also included in Table4.2: their locations makes them good indicators of the delay situation affecting JFK-ORD flights, and 296 aircraft flying from JFK to ORD flew into JFK from BOS in the two year period. Finally, we want to highlight the presence of the JFK-FLL departure delay in Table4.2. Fifty of the aircraft traveling from JFK to ORD came from FLL in 2007 and 2008, but this does not seem an strong argument to justify the presence of the JFK-FLL variable in Table4.2. Furthermore, the departure delay of flights leaving from JFK to FLL at time t would affect JFK-ORD departures around 6 hours later due to connectivity issues, and we have a 2 hour prediction horizon. Network effects seems the strongest reason why the JFK-FLL departure delay is the eight most significant OD pair in the network. All the other variables denote airports in the north east of the United States; however, as we saw in the case of the NAS state centroids, a high delay situation in a southern location (e.g. ATL) will typically affect the delay values in the north east due to network effects.

With respect to the number of OD delay variables to include in our models, we consider

Table 4.2: Influential OD pairs for JFK-ORD departure delay prediction.

Origin	**Destination**	**Delay Type**	**Variable Importance**
JFK	ORD	Departure	100
EWR	ORD	Departure	90.9
LGA	ORD	Departure	65.3
ORD	JFK	Departure	44
ORD	LGA	Departure	24.3
BOS	ORD	Departure	17
PHL	ORD	Departure	16.9
JFK	FLL	Departure	11.9
BUF	JFK	Arrival	11.4
LGA	ORD	Arrival	11

the top 10 variables, as we previously did for the airport variables. The average importance level for the 10th most important influential OD delay variable for the 100 most delayed OD pairs, which we will analyze in detail in Chapter 6 is only 16.5. More than 10 variables would not add much to our prediction models. As we mentioned previously, it is important to define a fixed number of variables because we want to compare different links' prediction models and changing number of variables would make the comparisons much less clear.

Table 4.3: JFK-ORD 2007-2008 aircraft rotations.

Previous Departure Airport	**Number of aircraft**
ORD	2,736
BUF	302
BOS	296
PWM	290
PIT	265
AUS	252
DCA	216
SYR	203
CLT	199
RDU	193

4.1.8 Other explanatory variables

In this section, we describe some additional explanatory variables that could be added to our delay prediction models, but which we do not include in the analysis presented in this thesis.

We have considered the most current delay state of airports and OD pairs, this is the delay state at the time the prediction is made. However, could the addition of past information improve or models? We could have variables depicting what happened (for example) 4 hours in the past. The NAS type-of-day is our only variable that goes back to the beginning of the day or even the previous day, but we do not look at specific OD pairs , airports or at NAS state information more than 2 hours in the past (size of the window of the moving median filter).

With the purpose of obtaining the NAS state variable, we clustered the delay state of the entire NAS, but we could instead cluster a certain area. For example, we could cluster the states of influential links identified. It may be possible to simplify our model by replacing a set of influential links delay variables with a categorical variable that models the most typical delay states of that set of links.

Chapter 5

Delay Prediction Models

In this chapter we test different classification- and regression-based delay prediction models. Our goal is to find the best prediction models, which we will used extensively in the remainder of this thesis. As we did in the previous chapter while analyzing the explanatory variables, we use the JFK-ORD departure link to evaluate the delay prediction models. The last part of this chapter studies the selected JFK-ORD departure delay prediction model in detail, through analysis of the errors variables importance, classification thresholds, and prediction horizons.

5.1 Training and test data sets

We first derive the training, and the test data sets needed to fit and test the performance of the different models we present in this chapter. We sampled 10 training sets (3,000 points each) and 10 test sets (1,000 points each) from the 2007-2008 data set. We fit and tested the prediction models for each of the 10 training and test set pairs. This allowed us to obtain a measure of the error variability and a good estimate of the test error. This methodology, called random sub-sampling or also Monte Carlo cross-validation (MCCV), consists on randomly partition the data into subsets, whose sizes are defined by the user. For each split, the model is fit to the training data, and the predictive accuracy is calculated using the corresponding split test data. The results are then averaged over the splits [15]. Random sub-sampling has been shown to be asymptotically consistent,resulting in more conservative predictions of the test data compared with cross-validation. The random-sub-sampling method gives a good estimate of the performance over external validation data [16].

The training and test sets created from the 2007-2008 data using over-sampling. Over-sampling is the deliberate selection of individuals of a rare type in order to obtain reasonably precise estimates of the properties of this type. In a population which includes a rare type, a random sample of the entire population might result in very few (or none) of these individuals being selected [14]. Over-sampling allows us to have a balanced data set, and to therefore avoid having more low delay data points in our training and test sets. This is especially important in the classification problem: If we want to classify future delays as high (e.g., over 60 min) or low (under 60 min), we want half of the points in our training and test sets to present delays of over (or under) 60 min. We used over-sampling in the regression problem as well, because it allowed us to compare classification and regression results, and evaluate the regression models over a rich data set (we do not want a large majority of data points in the data set to be low delay data points).

By applying over-sampling directly to the 2007-2008 data set, we still run into data issues. Due to the highly skewed delay distribution (see Figure 5-1), most of the low delay points (e.g., under 60 min delay) will have zero delay and belong to night time periods. To avoid this problem we eliminate from our data set all those data points where the output delay is zero. Figure 5-2 shows the histogram of the points eliminated for the JFK-ORD departure delay variable. We see that the vast majority correspond to night time periods. There is another important reason why delete these data points from our data set: As we described in Chapter 2 of this thesis, we work with delay state estimates, and interpolate delay information when there are no flights in the filtering window at a given time step. In doing so, we discovered that a significant number of delay prediction errors correspond to situations in which the estimated delay state variable value was zero, since there were no flights in the time window, but the explanatory variables depicted a high delay situation. In these situations, we correctly predict a high delay values, but the observed delay state value is zero. By deleting the zero delay data points, we avoid obtaining prediction performance values distorted by this data preprocessing issue. The histogram of the JFK-ORD departure delays for the 10 test data sets we use in our classification analysis (60 min threshold) is presented in Figure 5-3. This histogram presents a reasonable number of points for the different delay values, which will allow us to evaluate the performance of our prediction models.

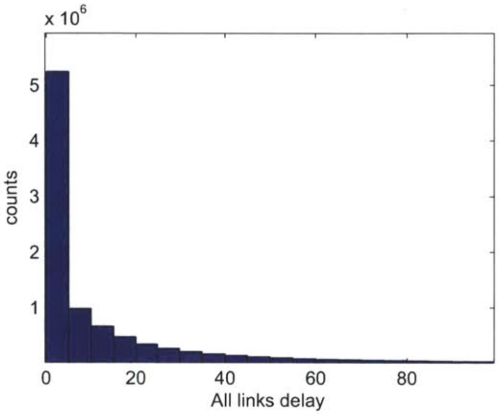


Figure 5-1: Histogram of the departure delay of all links in the simplified network.

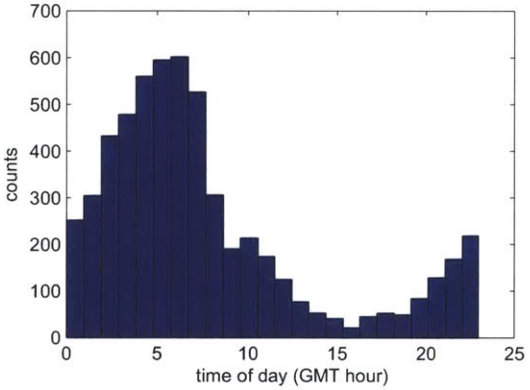


Figure 5-2: JFK-ORD zero departure delay data points by time of day.

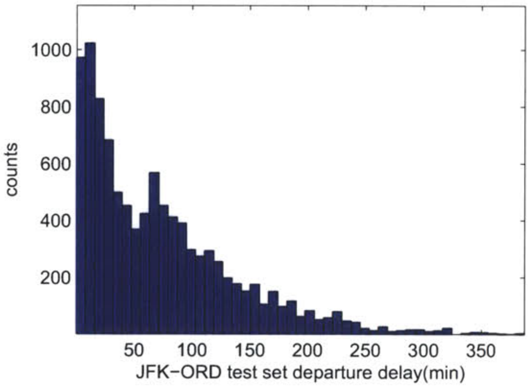


Figure 5-3: Histogram of the JFK-ORD departure delay 10 test sets data .

5.2 Collinearity analysis

Before starting the evaluation of the different JFK-ORD departure delay prediction models, we perform a collinearity analysis based on linear regression Variance Inflation Factor (VIF) values (categorical variables are treated as 0-1 variables). The VIF for the *jth* variable is calculated as follows:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{5.1}$$

where R is the coefficient of determination for the regression of the jth explanatory variable on the remaining $p-1$ explanatory variables. A large VIF value indicates that there is redundant information in the explanatory variables. Most authors consider values above 10 as high values, while others say 5. From the 61 explanatory variables in our model, after including the 0-1 dummy variables, five of the binary variables showed VIF values over 5. All these high values were associated with the NAS type of day and NAS previous type of day variables. These highest five VIF values are: 5.6, 6.2, 6.9, 8.2, and 9.3. In order to understand which variables lead to these VIF values, we eliminated different variables from the "rest of the variables set" ($p-1$ explanatory variables) in the VIF calculation and studied the effect on the resultant VIF values. We found out that for the NAS type-of-day variable and the NAS delay state variable, the airports and links delay state variables, and the NAS previous type-of-day were the variables leading to the highest VIF value of 9.3. By eliminating these 3 sets of variables from the NAS type-of-day, and NAS previous type-of-day VIF calculation, the largest VIF decreased to 5.5. Earlier in this thesis, Chapter 4, we saw that there is correlation between the NAS delay states and the NAS type-of-day, and that some NAS states only take place on certain types of days. There is also some correlation between the NAS type-of-day and the previous type-of-day, especially for high delay days when delay disruptions can last several days. Finally, the delay levels of different airports or links are correlated with the type-of-day, since the type-of-day information is obtained from clustering the link delays for the entire network and day. Overall, the VIF values obtained from including all the explanatory variables are not very high, and consequently we do not eliminate any of the variables. We instead use variable selection methods specific to each classification/regression model to eliminate variables if needed.

5.3 Classification models

In this section we evaluate the performance of different classification models in predicting the JFK-ORD departure delay. We consider a binary output, which indicates whether the output delay level is high or low. All the results presented below were obtained for a 60-min classification threshold and a 2-hour prediction window. Later in this chapter, the effect of changes in the classification threshold and prediction horizon on the performance of prediction models will be studied.

5.3.1 Classification based on logistic regression

Logistic regression is a generalized linear regression, which uses the logistic function as the link function. The following equation shows the structure of a typical logistic regression:

$$f(x) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}} \tag{5.2}$$

Logistic regression can not handle categorical variables directly, they need to be converted to binary variables. For this reason we have 61 variables instead of the original 26 variables; 41 binary variables, and 20 continuous variables. For each categorical variable we have the (number of categories minus one) binary variables (one category is the reference category).

The first step in fitting a logistic model was to identify influential data points. We first fit a logistic regression model with 3000 training observations, and then looked at Pearson residuals, deviance residuals, and leverage values. We eliminate all data points that satisfyed the conditions: $|Pearson\ residual| > 2$ or $|deviance\ residual| > 2$ or $leverage > 2 * mean(leverage)$. For this specific 3,000 point training data set, we identified 419 influential data points, on 14% of the data set. The next step was to fit a logistic model with the remaining 2,581 data points. The final misclassification error rate for the test data (1,000 points) was 22.9%. With respect to the variables' relevance in the logistic model, we saw than only 25 of the 61 variables had p-values under 0.05. Figure 5-4 shows the histogram of the p-values.

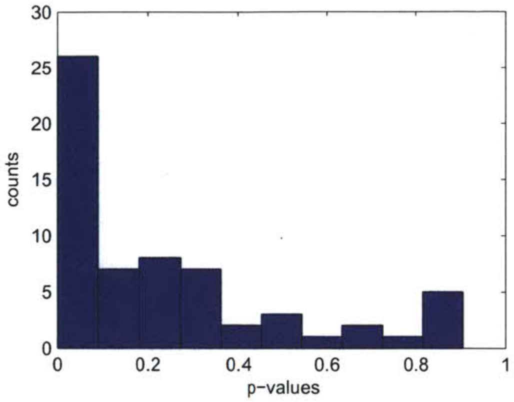


Figure 5-4: Histogram of the logistic regression model explanatory variables p-values.

These high p-values suggest that some of the variables are not needed in the logistic model. It therefore made sense to run a feature selection algorithm to eliminate the variables that do not play an important role in the logistic model. We used sequential forward selection and the logistic model deviance as the objective function to minimize termination tolerance: *chi2inv*$(.95, 1)$. Table 5.1 shows the order in which the variables were added, and their deviance values. Of the 61 initial variables only 25 were finally selected. Of the 6 original categorical variables, 3 were included in the model: time-of-day, NAS delay state, and NAS type-of-day. On the other hand, of the 20 delay states variables, 3 airport and 7 link variables were selected. The maximum p-value for this new model was 0.04: all the variables were significant at a 95% confidence level.

Table 5.1: Sequential Forward Selection results.

Iteration	Variable Added	Deviance
1-10	10 time-of-day categories: 4, 6, 8, 15, 16, 17, 18, 19, 21, 22 (eastern time)	4,159-2,975
11-12	NAS delay state: low, NYC high delay state	2,955-2,937
13-15	NAS type of day: NYC ORD medium delay, ATL high, NYC ORD medium high delay, NAS low delay.	2,922-2,895
16-18	Airports delay state: EWR departure delay, JFK departure delay, ORD arrival delay	2,886-2,859
19-25	Link delay states: BOS-ORD dep. delay, EWR-ORD dep. delay, JFK-FLL dep. delay, JFK-ORD dep. delay, LGA-ORD dep. delay, ORD-JFK dep. delay, PHL-ORD dep. delay	2,851-2,818

The test error value for this reduced model was 22.6%, only slightly lower than the value obtained for the model with all 61 explanatory variables. However, the lower number of variables makes this model easier to work with. Figure 5-5 shows the Receiver Operating Characteristic (ROC) curve, which shows how the TPR and the FPR varies as we change the decision threshold. The area under the ROC curve (AUC) is the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [17]. The AUC value is useful to compare different classifiers. For the logistic regression classifier, the AUC was 0.85.

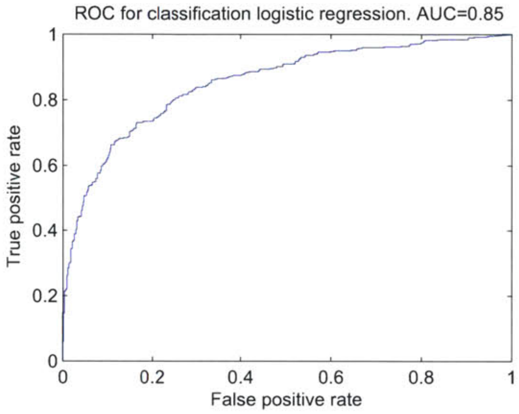


Figure 5-5: Logistic regression classification model ROC curve .

Finally, we obtained the logistic regression average test error and standard deviation, from 10 training, and test sets (MCCV). The values are: 23% test error, 1pp standard deviation. Later in this chapter, we use these values to compare the logistic regression model with other classification models.

5.3.2 Single classification tree

In this section, we study the performance of a single classification tree. Classification trees partition the input space into rectangles, and then fit a simple model in those rectangles. The steps are: Split the space into two regions by choosing an input variable and split point in order to achieve the best (LS) fit, and repeat this process again until a stopping condition is satisfied. The GINI index was used to grow the trees presented in this section; however, deviance was also tested with similar results.

A large tree will likely overfit, while a small tree will not capture enough detail. We would like to find the smallest tree that performs well, and to achieve this, we prune the tree using the misclassification rate as the performance criterion. Figure 5-6 shows the MCCV ($\pm\sigma$) error rate for different pruning levels (the higher the pruning level, the fewer nodes the tree has). We can see that for pruning levels around 15 to 30 the curve is almost flat, but for lower pruning levels the error rate increases (due to overfitting). For high pruning levels, the error also increases since we have trees with a very small number of nodes, which do not capture enough information from the training data. The minimum test error is 24.1%, and its standard deviation is 1.6pp. This implies that the logistic regression error rate calculated previously (23%) is lower than the single classification tree error rate. The classification tree AUC value is 0.79, which is also worse than that of logistic regression (0.85).

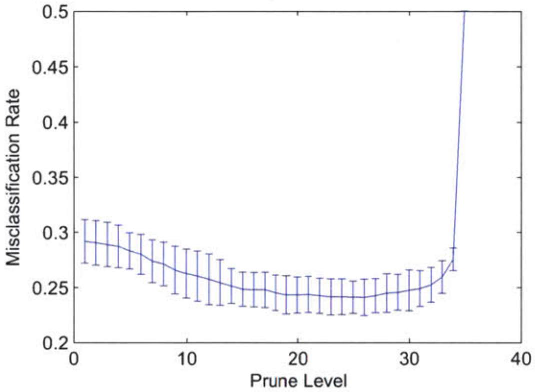


Figure 5-6: Single tree test error for different prune levels.

5.3.3 Ensemble of classification trees

The main idea behind ensemble classification methods is to aggregate predictions made by multiple classifiers in order to make a final prediction. In this section, we study three different ensemble methods: bagging, boosting and RF methods [13]. All three methodologies aggregate predictions from classification trees.

Bagging consists of sampling the training data set with replacement, and building a classifier (a tree in our case) for each bootstrap sample. Each tree is different and could provide a different prediction. The majority output is usually selected as the ensemble classifier output. By contrast, boosting combines the output of many weak" classifiers to produce a more powerful classifier. Each model (tree) is assigned a weight based on its performance on the training data. The final ensemble classifier output is the weighted sum of each individual classifier outputs. For each classifier, boosting follows an iterative procedure to adaptively change the distribution of training data by focusing more on previously misclassified records. Initially, all the training observations have the same weights, and after each iterations the weights of the misclassified observations are increased. Finally, RF is similar to bagging; however, the number of variables searched at each split is a random subset of the total variables set. RF works well for a large number of variables, and even when the number of variables is higher than the number of data points.

Before reviewing the results for each of the three ensemble methods, we describe a few specific configuration details. For boosting, we applied the AdaBoostM1 method, and a weak classifier where no fewer than 10% of the training data points could fall in a terminal node. For bagging and RF, we selected the minimum number of observations per tree leaf to be 1. In the RF model, the size of the random subset of variables searched at each split was chosen to be the square root of the total number of variables.

Figure 5-7 depicts the performance of the three methods for different numbers of trees on an specific training and test set for boosting, and the out-of-the-bag error for bagging and RF. We see that RF and bagging outperform boosting, but that RF is only slightly better than boosting. The minimum classification error (20.7%) was obtained for a RF with 91 trees. The MCCV test error value for the RF with 100 trees is 21.2% with a 1pp standard deviation, for bagging is slightly higher 22.1% with a 0.8pp standard deviation. The RF test error numbers are the best found so far, since the MCCV error was 23% for the logistic regression model, and 24.1% for the single classification tree model. Figure 5-8 shows the ROC curves for the RF, logistic regression and single tree models. We see here that the RF model performs better than the others, since it has the highest AUC value (AUC=0.87).

Based on these results, we chose the RF model as the best classification model, and the one used in the rest of this thesis. With the goal of having robust estimates of variable importance provided by the RF algorithm, we select 100 as the number of trees in the RF prediction models. As done in the logistic regression model, we could have eliminated

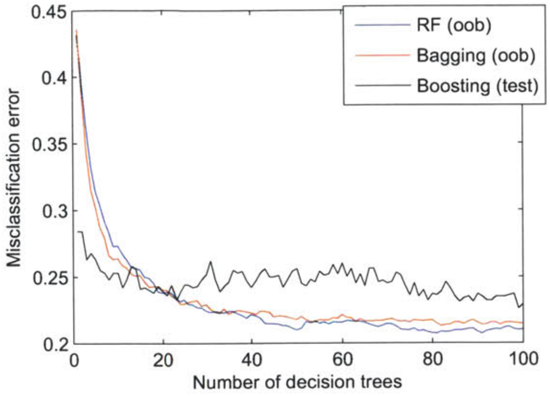


Figure 5-7: Ensemble methods performance for different number of trees.

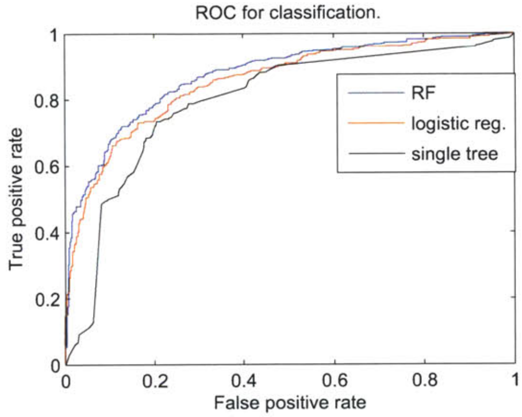


Figure 5-8: RF, logistic reg., and single tree ROC curve.

some variables from the RF model without significantly affecting prediction performance. However, RFs can handle a large number of variables, even if they do not provide much information or on the correlation among them. Having all the previously identified explanatory variables in this model will allow us to characterize different links by the importance of their explanatory variables, without having to carry an independent variable selection process for each link.

Finally, we calculate the test error value without over-sampling: data points are randomly sampled from the entire data set. The mean MCCV test error is 18.4% with a 1pp standard deviation. As expected, the MCCV mean test errors are lower without over-sampling since there are more easy-to-predict low delay data points.

5.4 Regression models

In this section, we study the prediction performance of different regression models for the JFK-ORD link. The output of the prediction model is continuous, and represents the predicted delay level in minutes. The results shown here were obtained for a 2-hour prediction horizon, using over-sampling (60 minute threshold) to generate the data sets. The mean and median absolute errors are used to evaluate the regression models performance.

5.4.1 Linear regression

In linear regression, a linear function is used to define the relationship between the explanatory variables and the output variable. As in the case of logistic regression, categorical variables are first converted into binary variables, resulting in 61 variables. Before fitting a linear model, we perform a diagnostic analysis and eliminate from the training set all data points with the following characteristics: Studentized residuals larger than 3, leverage values greater than $3p/n$ (p is the number of variables, n the number of points in the training data set), dffits values greater than $3\frac{\sqrt{p}}{\sqrt{n}}$, and dfbetas greater than $\frac{3}{\sqrt{n}}$. We calculate the MCCV test errors for a linear model with all the variables, and a reduced linear model with only 20 variables. This reduced model is obtained with the same forward feature selection procedure used in logistic regression, with the R-squared value as the objective function that is maximized (termination tolerance 0.001). Table 5.2 presents the MCCV performance numbers for both models, showing that the reduced model outperforms the complete model to a small extent.

Table 5.2: Linear Regression models performance

LR Model	Median error (min)	Median error std (min)	Mean error (min)	Mean error std (min)
20 variables	23.7	0.9	34.2	0.9
All variables	23.8	1	35.2	1

5.4.2 Single regression tree

The main difference between regression trees and classification trees is that terminal nodes have specific delay values associated instead of binary values. As was done for the classification trees, the GINI index was used as the splitting criterion to grow the regression trees. Figure 5-9 shows the MCCV mean and median test error and standard deviations for different pruning levels. The minimum average median test error is 26.2 min (std=1.4 min), this value is reached for a pruning level of 436, and the minimum mean error is 38.5 min (std=1.4min), reached for the pruning level of 507. Consequently, the linear regression model presented in the previous section performs better than this single regression tree.

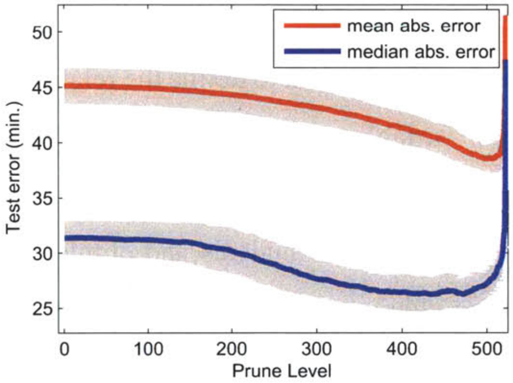


Figure 5-9: Mean and median test error for different pruning levels ($\pm\sigma$).

5.4.3 Ensemble of regression trees

In this section, we evaluate the same three ensemble methods that we evaluated in the Section 5.3.3 for classification. We use the same parameters that were previously used for bagging and RF methods, and in boosting we use the LSBoost method (the AdaBoostM1 method is only for classification).

In Figure 5-10, we see the mean and median absolute errors for specific training and test sets, and different number of trees. Boosting performs significantly worse than the other two methods, and there is no significant difference between bagging and RF. With respect to the number of trees, we see that for more than 30 trees, the error curves are almost flat; the extra trees do not improve the prediction model performance. However, it is interesting to have more than 30 trees in our model to obtain a more robust estimate of the importance of the different variables.

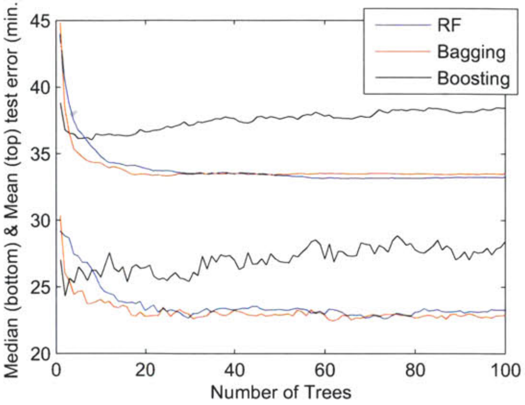


Figure 5-10: Ensemble methods median and mean test error for different pruning levels.

The RF model MCCV median and mean test errors are 24.6 and 33.6 min, and for the bagging model, 24.8 and 34.6 min respectively. These values are slightly higher than those obtained from the linear regression model. However, we choose RF as the best model because of the relative variable importance values that it provides. As we will see in the rest of this chapter and in Chapter 6 the explanatory variables' importance values are key in our analysis, and RF provides good estimates because it randomly selects subsets of variables, and does not only focus on the best variables to generate the tree splits (as bagging does).

Finally, as we did in the classification problem, we calculate the test error numbers without over-sampling. The MCCV median error is 19.8 min, and the mean error is 28.7 min. These test error numbers are lower than those obtained with over-sampling because we randomly sample the training and test data sets, and therefore, there are more a larger number of easy-to-predict low delay data points in the data sets.

5.5 Detailed analysis of random forest model

The RF model was selected from among the delay prediction models studied. In this section, we perform a detailed analysis of: the importance of the different explanatory variables in the RF model, the data points for which the model does not work well (and the reasons why), and the effect of the classification threshold and prediction horizon on model performance.

5.5.1 Importance of explanatory variables

Our goal in this section is to identify the most relevant explanatory variables in the JFK-ORD departure delay prediction model, and to find similarities and differences between the explanatory variables importance in the classification and regression problems.

The explanatory variables' importances were obtained as the average of the RF variables' importance for each of the 10 training data sets. Table 5.3 shows the results for classification and regression (using a 2h prediction horizon, and 60 min over-sampling threshold). We first analyze the importance values of the classification variables. Time-of-day is the second most relevant variable, only the JFK-ORD departure delay exceeds its importance. None of the airport delay variables manifested as relevant prediction variables; ORD departure delay has the highest importance value (36.4). With respect to the links' delay variables, the JFK-ORD departure delay has the highest importance of all variables (100). The next two most important links are also departures from NYC to ORD: EWR-ORD and LGA-ORD. The results of the regression problem present some interesting differences. The time-of-day is less relevant with only a 59.9 importance level. This indicates that the time-of-day is a very good prediction variable when we are only interested in knowing whether delays are high or low (the classification problem); however, if we are looking for more detail (the regression problem) the time-of-day is not a good prediction variable. The NAS delay state variables is significantly more important in regression (89.4) than classification (44.4). A possible reason for this behavior is that when we are in a NAS delay state in which either JFK or ORD delays are high, the JFK-ORD departure delay will likely be within a certain high-delay interval; let us say between 50 and 70 minutes. This can be valuable information if we are trying to estimate the exact value of the delay (regression), but if we want to decide if the delay level is above or below 60 minutes (classification), this information is less relevant. Airport delay variables are even less important in regression than in classification. Finally, we see that the same three links are the most important links in regression, but the gap between the most important link (JFK-ORD) importance level and the second most important link is higher. This fact, together with the lower airport importance levels indicates that for JFK-ORD delay prediction, the delay states of the defined airport and link influence areas are more important in classification than in regression.

Table 5.3: Classification and Regression variables importance for the JFK-ORD departure delay prediction model

Explanatory Variables	**Variables Classification Importance**	**Variables Regression Importance**
Time-of-day	89.8	59.9
Day-of-week	7.3	3.1
Season	3.6	1.2
NAS delay state	44.4	89.4
NAS type-of-day	38.1	22.8
NAS prev. type-of-day	21.9	11.4
BOS departure delay	29.6	13.4
DCA departure delay	31.6	17.1
EWR departure delay	28.4	12.8
JFK departure delay	33.3	20.7
LGA departure delay	29.1	14.1
ORD departure delay	36.4	15.6
PHL departure delay	24.4	12.3
DCA arrival delay	25.2	13
JFK arrival delay	30.1	18.5
ORD arrival delay	34.7	12.9
BOS-ORD dep. delay	36.1	19
EWR-ORD dep. delay	79.4	62.3
JFK-FLL dep. delay	29.2	16.7
JFK-ORD dep. delay	100	100
LGA-ORD dep. delay	59.8	52.2
ORD-JFK dep. delay	58.5	40.6
ORD-LGA dep. delay	31.7	21.9
PHL-ORD dep. delay	36.6	22.8
BUF-JFK arr. delay	20.7	15.7
LGA-ORD arr. delay	23.8	13.3

5.5.2 Analysis of errors

In this section, we conduct an exhaustive error analysis for the RF classification model; some notes about the regression model will are included. The goal is to identify situations where our model does not perform well, and explain why. The analysis of the errors can help us identify ways to improve the prediction models' performance.

We first consider datapoints misclassified by the JFK-ORD prediction model misclassified data points for a 60 minute classification threshold and a 2 hour prediction window. The misclassified points are those obtained from the 10 test data sets. Figure 5-11 shows the future JFK-ORD departure delay for the misclassified data: our model classified all points over 60 min as low delay points (negative class), and vice versa. We can see that there is a high concentration of data points around 60 min; this means that when the future delay value gets closer to 60 min it is harder to classify that data as high or low delay data. It is also noticeable that between 10 and 50 minutes, we have an approximately constant number of misclassified points. Low delay points are easier to classify, but we have a larger number of low delay points in the test sets, and this increases the number of misclassified low delay data points. Figure 5-12 shows the normalized version of Figure 5-11. As we expected, the misclassification error rate is maximum for 60 minutes delay, and as we move away from this point, the error rate decreases. For high delay values we do not have many points (see Figure 5-3). As a results, the results are not reliable and we can go from a 0% misclassification rate to 65%.

Next, we analyze dependence of the misclassification rate on time of day (Figure 5-13). When the day starts the error rate is low, (slightly over 10%), it increases around 9am to 17%, and it stays at this level for most of the day (9am 10pm). It is importance to notice that the model achieves an steady error rate for different JFK-ORD delay levels, since delay increases as the day progresses (see Figure 4-2). However, at the end of the day the error rate increases significantly. The reason is the lack of data depicting the delay states of the different elements in the system. The error rate rises for predictions after midnight, because at that time there is not enough flight data to have a reliable delay state value of the different links and airports our model depends on. We also consider the variations of the FN and FP

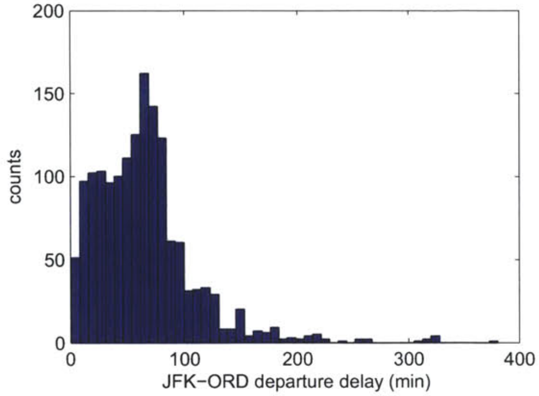


Figure 5-11: Histogram of the future JFK-ORD departure delay values for the missclassified data points.

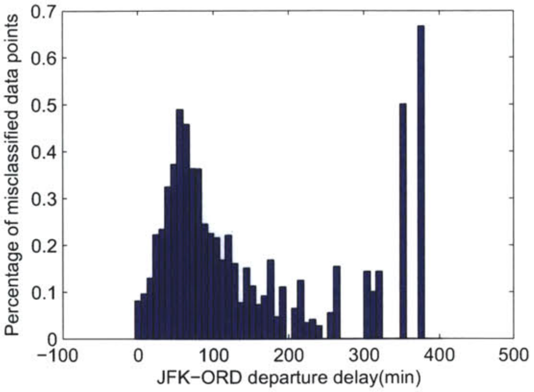


Figure 5-12: Missclassification rate vs future JFK-ORD departure delay .

rates with the time of day. In Figure 5-14, we can clearly see that FNs are dominant at the beginning of the the day, and as the day progresses, FPs become dominant. The reason for this behavior is two fold: On the one hand, our model relies on the network state, and consequently it is hard to detect when delays rise or drop significantly; on the other hand, we have the effect of the time-of-day variable. Early in the day, delays tend to be low. If there is no indicator of a future high delay situation at the time, we will likely predict low delay, even though delays could be high (a storm could hit the airspace in 2 hours) leading to a high FN rate. The reverse situation takes place late in the day. If the network delay is high, we likely to predict high delay; however, there is a point when delays could drop due to the lower traffic volume late in the day. This leads to a high FP rate. The second factor we depicted was the time of day variable. This is a very relevant variable in our model with an 89% importance level. Typically, delays are low in the morning and high at the end of the day. The time of day variable will push delay predictions down early in the day (leading to FNs), and it will push them up late in the day (leading to FPs).

A different view of the error rate and FP/FN rates is presented in Figure 5-15, where the error rate for different values of the NAS delay state at the time of prediction is presented. The higher error rate is reached for the medium delay level (State 1). This NAS state is typically a transition state to a higher or lower delay situation, and as we mentioned previously, there may not be enough evidence in the network of the delay rise or drop at

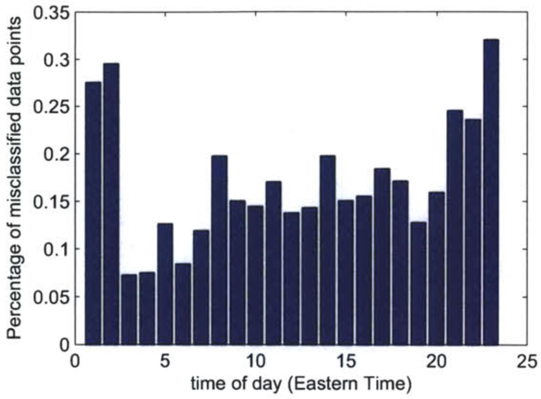


Figure 5-13: Misclassification rate versus time of day.

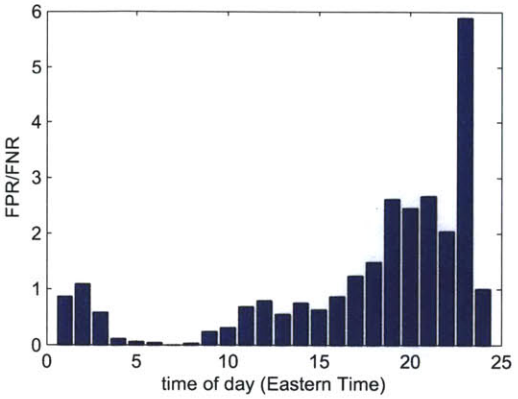


Figure 5-14: FPR/FNR versus time of day .

the prediction time. It is worth noting the high error rate of the ATL high delay state (State 6) compared to the other high delay states (States 2, 3 or 5). The reason can be that in this case, the main delay source is not JFK or ORD, leading to a more unpredictable situation. On the other hand, when we are at a high NAS delay state FPs dominate over FNs; it is difficult to know when the high delay situation is going to end. For the low delay state (State 4), FNs are clearly dominant; it is difficult to detect when delays go from low to high when there are not enough signs of the change in the delay situation at the time of prediction. We also studied the misclassification error rate for different months of the year, but we did not see any remarkable differences.

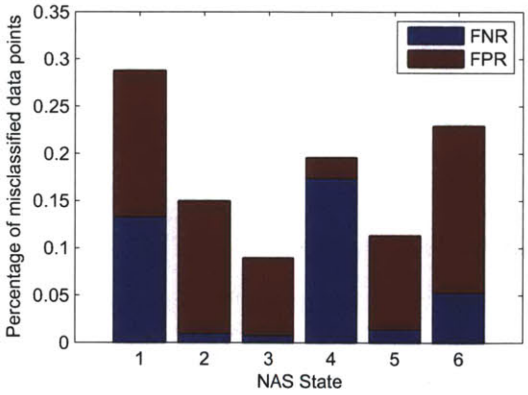


Figure 5-15: Misclassification error by NAS state.

Next, a detailed study of two misclassified data points is presented. Table 5.4 shows the two data points' explanatory variables' values. For the first of the data points, we predicted low delay (under 60 minutes), and the future departure delay was 306 minutes. The RF values for this data point, which indicate the prediction probability of belonging to each of the classes, were 85.8 (for the low delay class), and 14.2 (for the high delay class). The three most important variables value were (see Table 5.4): 23.6 min for the JFK-ORD departure delay, 0 min for the EWR-ORD departure delay, and 11 am for the time-of-day (typically high delays take place later in the day). The cause for the 306 minute delay could have been a mechanical issue, or any other local circumstance that does not have an effect in the delay state of the network, and consequently our model does not capture. It could also be the beginning of a high network delay period; however, that day was not classified as a high delay day. It is important to notice that the proposed model is a network-based model. The delay prediction is based on the delay state of the different network elements at time t. Misclassified data points points show situations in which either local or individual flight issues take place without having an effect on other elements of the network, or situations where there are not enough signs of the network delay increase or decrease at the time of prediction.

For the second misclassified data point, we have the opposite situation: We predicted the JFK-ORD departure delay to be high (over 60 min) and the actual delay was only 6 min. In this case, there are many signs of a high delay situation in the explanatory variables (see Table 5.4); for example, we are in the Chicago high delay NAS state, and the JFK-ORD delay is 204 min. We made this delay prediction at 10pm (Eastern Time), meaning that we were predicting midnight delays. At midnight, the demand level drops, which causes congestion to also drop. Aircraft accumulate large delays during the day, and these high delays can be carried until the end of the day. Nevertheless, aircraft that were not affected by those delays or had enough turnaround time buffer to absorb previously accumulated delay could depart on time when congestion drops. Once again, this is a difficult situation for our model, since it depends on the delay state of the network. The only variable that can help in this type of situation is the time-of-day, because it pushes delays down late at night, when delays typically decrease.

The different findings presented above for the classification model are applicable to the regression model. As an example, in Figure 5-16 we have the median prediction error

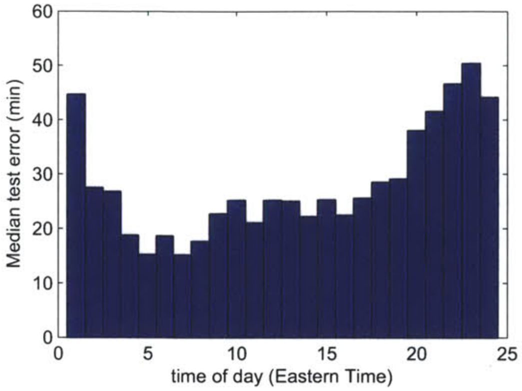


Figure 5-16: Regression median test error by time of day.

versus the time of day, and we can see that the error varies with time-of-day the same way that the misclassification error did (Figure 5-13)

5.5.3 Effect of classification threshold

The classification threshold value affects our prediction models performance. In this section, we evaluate the impact of changes in the classification threshold on the test error, and the effect on the importance of the different variables in the model.

The comparative analysis was performed for 45, 60 and 90 min classification thresholds and a 2 hour prediction horizon. With respect to the test errors, we obtained the following results: 23.13% (std=1.5) for the 45 min threshold, 21.2% (std=1%) for the 60 min threshold, and 19.39% (std=1.1%) for the 90 min threshold. The test error decreases as the classification threshold increases, since there are clearer indications of whether the future delay will exceed 90 min, than whether it would exceed 45 min.

With respect to the variables' importance value, the time-of-day variable importance decreases as the classification threshold increases (see Table 5.5). The time-of-day variable has difficulties explaining high delay values which do not take place often: For high threshold values, we need high delay signs from different elements of the network. Another variable with a higher importance value for the lowest classification threshold is the NAS type-of-day. The JFK-ORD departure delay distributions for the different type-of-day clusters are less centered around 45 minutes than around 60 or 90 minutes, allowing to better distinguish between delays over and under the threshold. The higher importance of the ORD departure delay for the 45 min threshold is also worth noting. Finally, the OD pairs delay variables show that as the threshold increases, the third most important variable's importance (LGA-ORD departure delay) decreases significantly, being JFK-ORD departure delay, and EWR-ORD departure delay the only two variables with an importance level over 50 for the 90 min threshold. This means that the correlation between the LGA-ORD departure delay variable and the future JFK-ORD departure delay decreases as the level of delay of JFK-ORD departures increases, and the EWR-ORD departure delay variable keeps a similar degree of correlation for different levels of delay.

5.5.4 Effect of prediction horizon

In this section, we study changes in the length of the prediction horizon. As done earlier, we evaluate the prediction model's performance, and the most significant changes in the explanatory variables' importances.

Table 5.6 depicts classification and regression prediction performance for four different prediction horizons: 2, 4, 6, and 24 hours. We see that the model performance does not decrease much as we increase the prediction horizon length: from 2 to 6h the classification error only increases 3.9pp.

Tables 5.7 and 5.8 show the most interesting changes identified in the explanatory variables importance value. The time-of-day variable for both classification, and regression plays a more important role as the prediction horizon increases. The delay state of the different elements in the network decrease prediction power as we increase the time horizon; however, the time-of-day variable is not affected by the time horizon leading to a higher importance value for the time-of-day variable. It is important to notice the high importance value of the NAS type-of-day categorical variable for a 4 and 6 hour prediction horizon. For a 24 hour prediction horizon the type-of-day variables importance drops, since in this case it is associated with the previous day, and not the day in which we are making the prediction. Finally, as a representative example of the reduction of the links delay variables importance as we increase the time horizon, we have the JFK-ORD departure delay. We see that this variable goes from being the most important one for a 2h horizon to have an importance level less than 30 for a 24h horizon, for both classification and regression.

Table 5.4: Sample 1 and Sample 2 explanatory variables value

Explanatory Variables	**Variables Importance**	**Sample 1 (min)**	**Sample 2 (min)**
Time-of-day	89.8	11 am eastern time	10pm eastern time
Day-of-week	7.3	Tuesday	Sunday
Season	3.6	Medium delay	Medium delay
NAS delay state	44.4	NAS medium delay	Chicago high delay
NAS type-of-day	38.1	NYC, ORD medium delay	Chicago high delay
NAS prev. type-of-day	21.9	NYC, ORD medium delay	Atlanta high delay
BOS departure delay	29.6	8.3	25
DCA departure delay	31.6	7.1	111
EWR departure delay	28.4	7.3	95.3
JFK departure delay	33.3	14.2	41
LGA departure delay	29.1	18.8	73.5
ORD departure delay	36.4	13.7	101.4
PHL departure delay	24.4	7	58
DCA arrival delay	25.2	5.5	33.3
JFK arrival delay	30.1	7.9	42.9
ORD arrival delay	34.7	7.5	113.5
BOS-ORD dep. delay	36.1	4.5	148
EWR-ORD dep. delay	79.4	0	133.7
JFK-FLL dep. delay	29.2	18.8	28.9
JFK-ORD dep. delay	100	23.6	204.3
LGA-ORD dep. delay	59.8	20.1	119.3
ORD-JFK dep. delay	58.5	23.3	77.9
ORD-LGA dep. delay	31.7	21.7	70.7
PHL-ORD dep. delay	36.6	1.3	161.3
BUF-JFK arr. delay	20.7	0	0
LGA-ORD arr. delay	23.8	0	134.6

Table 5.5: Explanatory variables' importances for JFK-ORD for different classification thresholds

Explanatory Variables	**Variables Imp. (th=45min)**	**Variables Imp. (th=60min)**	**Variables Imp. (th=90min)**
Time-of-day	100	89.8	61.7
Day-of-week	8	7.3	5.1
Season	4.3	3.6	2.5
NAS delay state	21.6	44.4	21.8
NAS type-of-day	64.5	38.1	45.9
NAS prev. type-of-day	26	21.9	17.2
BOS departure delay	32.2	29.6	19.3
DCA departure delay	32.7	31.6	18.4
EWR departure delay	36.5	28.4	20.4
JFK departure delay	48.9	33.3	36.9
LGA departure delay	31	29.1	19.7
ORD departure delay	60.8	36.4	32.3
PHL departure delay	31.9	24.4	19.7
DCA arrival delay	33.6	25.1	19.8
JFK arrival delay	36.8	30.1	26.2
ORD arrival delay	42.9	34.7	41.9
BOS-ORD dep. delay	43.5	36.1	22.4
EWR-ORD dep. delay	89.3	79.4	87.5
JFK-FLL dep. delay	34.5	29.2	17.3
JFK-ORD dep. delay	95.7	100	100
LGA-ORD dep. delay	70.3	59.8	49.8
ORD-JFK dep. delay	66.4	58.5	49.8
ORD-LGA dep. delay	30.1	31.7	20.6
PHL-ORD dep. delay	42.4	36.6	44.4
BUF-JFK arr. delay	23.3	20.7	19.5
LGA-ORD arr. delay	20.9	23.8	18.1

Table 5.6: Prediction horizon analysis.

Prediction horizon (h)	**Class. test error (%)**	**Class. test error std (pp)**	**Reg. median error (min)**	**Reg. median error std (min)**
2	21.2	1	24.49	0.7
4	23.06	1.2	27.45	0.8
6	25.12	1.5.	29.52	0.93
24	32.23	1.4	32.8	1.59

Table 5.7: Effect of prediction horizon on explanatory variables importance for classification.

Variables	Horizon (2h)	Horizon (4h)	Horizon (6h)	Horizon (24h)
Time-of-day	91.6	99.1	100	100
NAS type of day	38.1	61	62.7	17.9
JFK-ORD dep. delay	100	39.8	19.9	25.9

Table 5.8: Effect of prediction horizon on explanatory variables importance for regression.

Variables	Horizon (2h)	Horizon (4h)	Horizon (6h)	Horizon (24h)
Time-of-day	61.3	100	100	100
NAS type of day	20.7	73.8	75.8	13.2
JFK-ORD dep. delay	100	36.5	15.8	18.2

Chapter 6

Analysis of the 100 Most-Delay OD Pairs

In this chapter, we analyze the performance of the RF prediction model presented in the previous chapter for the 100 most delayed OD pairs in the 2007-2008 data set. We selected the 100 OD pairs with the highest delay in order to avoid a shortage of high delay data points. Figure 6-1 depicts the 100 OD pairs considered in the analysis.

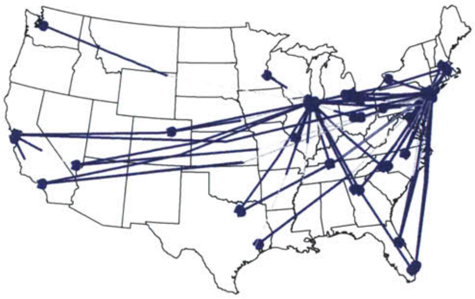


Figure 6-1: 100 most-delayed OD pairs.

The goal of this chapter is to study prediction models for the different OD pairs and to identify sets with similar characteristics (e.g., similar prediction performance, or explanatory variables' importance). In the analysis presented hereafter, we evaluate the potential of the different explanatory variables included in the prediction model. In the JFK-ORD prediction model which we discussed in the previous chapter, some of the variables were found not to be relevant, but they help us here to better understand the characteristics of different links. For example, the NAS delay state variable may be the most significant prediction variable for a particular link, and one of the least significant variables for another one; in this scenario, we could conclude that delays on the first link have a high correlation with the global NAS delay state.

Another important objective of this chapter is to validate the delay prediction model on different OD pairs. The results presented in the previous chapter were obtained for the JFK-ORD OD pair, and we do not know if these results are specific for this link or they can be generalized to any OD pair in the network. As we described in the explanatory variable selection process (Section 4.1), our goal is to develop a prediction model that can be used for any link in the network.

The majority of the results presented in this chapter were obtained with a 2h prediction horizon, and a 60 minute classification threshold. Some interesting results for different time horizon lengths and classification thresholds are also presented. Both classification, and regression models are studied in this chapter, with a greater focus on the classification models.

6.1 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.

6.1.1 Classification performance

Figure 6-2 shows the test error histogram for the 100 most delayed OD pairs. The test error ranges from 11.3% to 28.8%, with an average value of 19.1%. The link with the lowest test error is EWR-ATL (11.3%), and the one with the highest is LAS-SFO (28.8%). Delays for flights arriving or departing from SFO are difficult to predict: The average test error rate for links that have SFO as either the origin or the destination is 23.3%. We find that 90% of the analyzed links have a test error standard deviation that is less than under 1.7 percentage points. The empirical cdf of the test error standard deviations of all the links is presented in Figure 6-3.

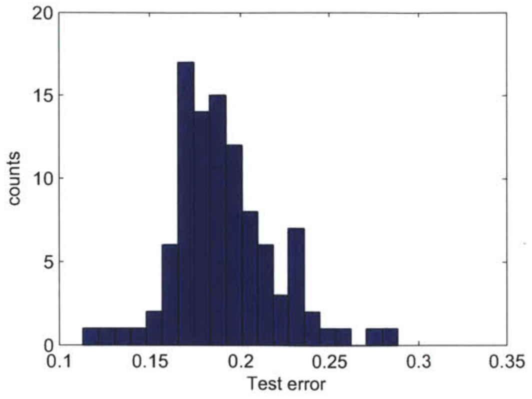


Figure 6-2: Classification test error histogram for the 100 most-delayed OD pairs.

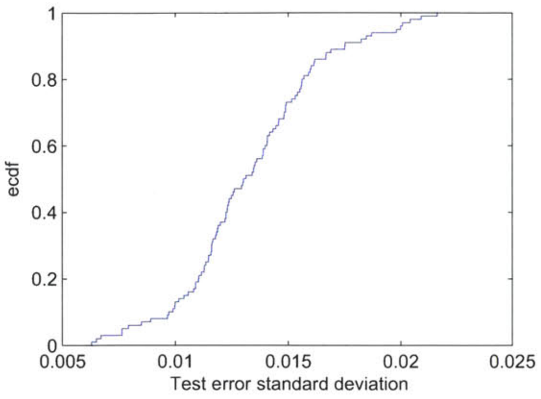


Figure 6-3: Empirical cdf of the standard deviation of the classification test error for the 100 most-delayed OD pairs.

If we break down the test error into false positive and false negative error rates (FPR and FNR, respectively) we find that the FNR is clearly dominant. For the 100 most delayed links, the average FNR is 23.62%, while the average FPR is 14.6%. Additionally the FNR rate is higher than the FPR for all OD pairs. In other words, the classifier is more likely to miss a high delay link than it is likely to predict high delay when actual delay on the OD pair is low. This behavior is because our prediction model considers the delay states of the different elements in the network, and therefore does not accurately capture local delay causes (such as mechanical issues). If delays in the relevant network elements are high, we will likely have a high delay situation in two hours later in our link of interest; however, if the network delay is low, we could still have a high delay in two hours later due to a local issue that affects only a certain flight. Figure 6-4 shows the FPR and FNR versus the test error for all the studied OD pairs. We note that the separation among the FP points and FN points increases as the test error increases. For the lowest test error OD pair, FNR/FPR ratio is 1.3, while for the highest test error FNR/FPR=1.9, showing that the FNR dominance increases with the test error. In the OD pair with the highest test error (LAS-SFO), the prediction model misclassifies high delay points almost twice as often as the low delay points.

Table 6.1 summarizes the most significant correlation results between the test error values and the explanatory variables importance. The three temporal variables show a clear

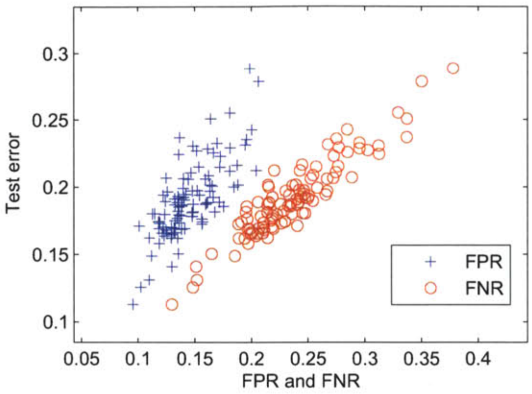


Figure 6-4: FNR and FPR scatter plot for the 100 most-delayed OD pairs.

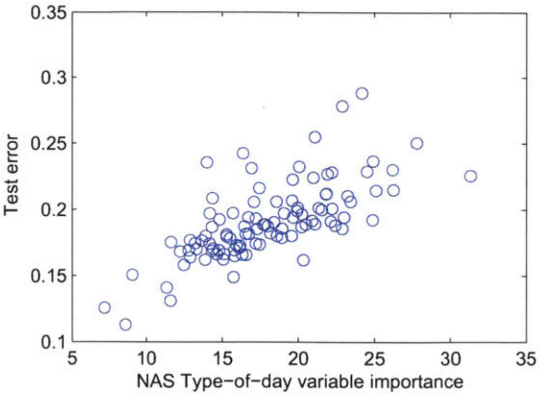


Figure 6-5: Test error versus type-of-day importance for the 100 most-delayed OD pairs.

positive correlation with the test error, specially the day-of-week and season variables. These results show that the temporal variables tend to have high importance 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 high correlation of the test error with the previous type-of-day variable (0.7) is worth noting. This means that when delays are hard to predict, the information on the previous day becomes more important. Figure 6-5 depicts the correlation of the test error with the type-of-day importance (0.43). 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 signs of the correlation coefficients. The airports' importance is positively correlated with the test error, and the OD pairs importance is negatively correlated. This means that having a high airport importance is typically a sign of poor prediction performance, perhaps because it suggests that none of the OD pairs are good predictors.

6.1.2 Regression performance

Next, we study the regression problem, and compare its performance with the results obtained for classification. We use the same data set used in Section 6.1.1.

Figure 6-6 shows the histogram of the median test error for the 100 links studied. The median error values range from 15.6 min (EWR-ATL) to 36.4 min (LAX-HNL), and the average median test error is 20.9 min. As we can see in Figure 6-7 the standard deviation

Table 6.1: Correlation between the test error and explanatory variables importance.

Explanatory Variable	**Correlation with classification test error**
Time-of-day	0.43
Day-of-week	0.74
Season	0.71
Previous type-of-day	0.7
Airports delay (max. importance)	0.33
OD pairs delay (max. importance)	-0.16

of these error values is low, and the 90th percentile of the distribution is 1.17 min. The gap between the highest median error value (LAX-HNL) and the second highest (SFO-JFK)in Figure6-6 is worth noting. Since neither of these links had the highest test error in classification, we can ask the question, do links with high classification test error also have high regression test error? To answer this question, we plot classification error versus regression error (Figure 6-8). Although there is a strong positive correlation (0.78), some specific links perform significantly differently in the classification and regression problems. The highlighted data point in Figure 6-8 corresponds to the CLT-LGA departure delay prediction model. The classification test error for this link is 22.6%, which is high and in the 87th percentile of the classification error distribution, but the regression median test error is only 20.2 minutes, which is in the 40th percentile of the regression error distribution. This shows that a good performance in the regression problem does not necessarily mean good performance in the classification problem, and vice versa. The problems are different: in the classification problem we need information to allow us differentiate between high and low delay (relative to a given threshold), but in the regression problem we need information to predict the *value* of the future delay. For a specific link, it may be easier to predict whether or not the future delay will be over 60 min. than to predict its exact delay value.

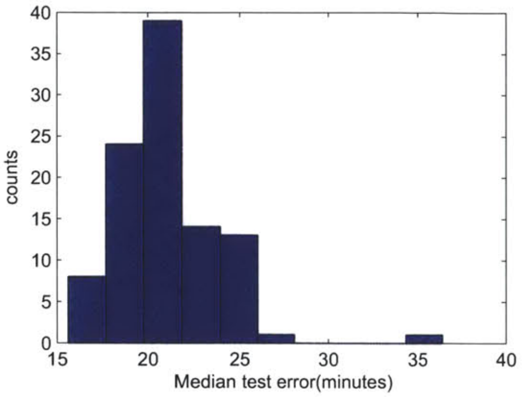


Figure 6-6: Median test error histogram for regression.

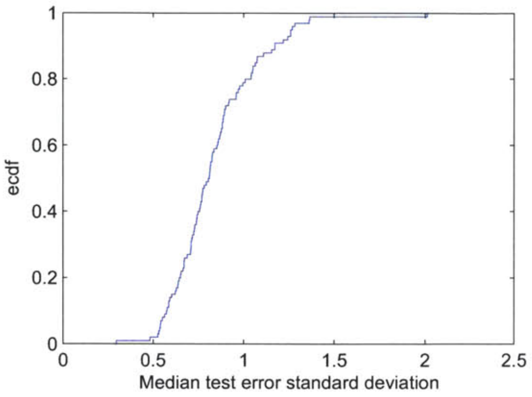


Figure 6-7: Empirical cdf of the standard deviation of the median test error for the regression problem.

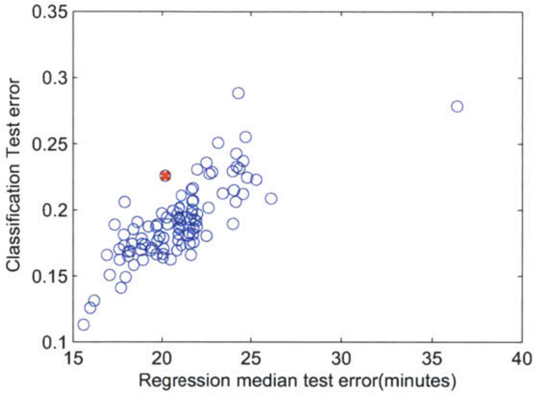


Figure 6-8: Classification vs. regression test error.

6.2 Clustering of delay prediction models

The delay prediction models that we have proposed include a large number of variables. Some of them are not relevant for certain OD pairs, but are relevant for others. The importance values given by the Random Forest algorithm will allow us to characterize the different OD pairs according to the importance of the different explanatory variables. Our goal in this section is to aggregate the 100 most delayed OD pairs into clusters, whose members share prediction models with similar characteristics.

The first step is to define a set of clustering variables, that reflect the most significant characteristics of the prediction models. We would like to cluster the prediction models according to their performance (test error), and the importance of the different explanatory variables. We choose as the clustering variables' the test error and the following variables importance values: time-of-day, day-of-week, season, NAS delay state, type-of-day, previous type-of-day, departure's airport departure delay, arrival airport's arrival delay, OD pair's departure delay, OD pair's arrival delay, the three airports with the highest importance value in the influence area, and the three links with the highest importance value in the influence area.

Once the clustering variables are defined we need to choose the number of clusters. As we did previously, we look at the k-means total within cluster distances for different number of clusters. Figure 6-9 shows the value of the total within-cluster distances. We can see that the total distance drops significantly until five clusters, after which the curve is flatter. For this reason, we choose five as the number of clusters.

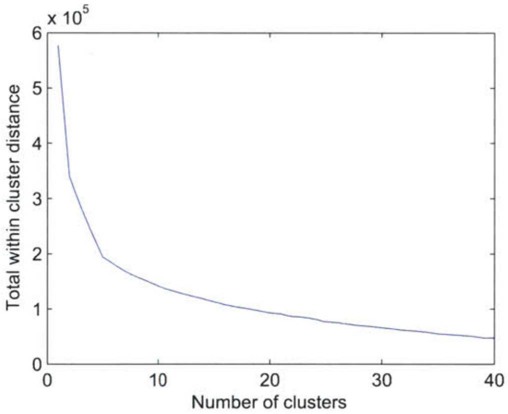


Figure 6-9: Total within cluster distance for different number of clusters.

Table 6.2 shows the values of the clustering variables for each of the 5 clusters' centroids, and Figure 6-10 depicts the OD pairs belonging to each cluster. Cluster 1 is characterized by a low test error (18.6%), high departure airport departure delay importance (98.4), low arrival airport arrival delay importance (0.9), and low importance for all the links' delay state variables (highest value 56.3). The links belonging to Cluster 1 have a large, highly congestion airport as the origin (JFK, EWR, LGA, ORD, PHL), and a smaller and/or less congested airport as the destination (see Figure 6-10). Cluster 2 has the highest test error value (20.6%), and highest time-of-day importance (100). The links belonging to this cluster are those whose delay is harder to predict, because none of the airports or links delay state variables are good predictors. Consequently, the time-of-day plays the most important role in the model. We can see that in Cluster 2, both the highest airport importance (44.4) and highest link importance (62.7) are close to their lowest values for all clusters. In Cluster 3, both the departure airport departure delay and the prediction OD pair departure delay have high importance values (79.7 and 82.3 respectively). In Figure 6-10 shows that as in Cluster 1, Cluster 3 links also have one of the major airports as the departure airport. The difference is that in Cluster 1, the departure delay of the prediction link was not relevant, while in Cluster 3, both the departure airport departure delay and prediction link's departure delay play important roles. Cluster 4 is the smallest of the clusters with only seven elements, and it is characterized by the highest arrival airport arrival delay importance (87.4), and a high prediction link departure delay importance (86.8). Five of the 7 links are LGA arrivals, and the remaining two are ORD arrivals (see Figure 6-10). Finally, Cluster 5, the cluster with the largest number of data points, is characterized by a high importance value for the prediction link's departure delay (95.5), and also high importance of the two most important links' delay state variables (100 and 71.6, respectively), but low airport variables' importance values (highest importance: 38). OD pairs belonging to this cluster arrive at one of the large and heavily-congested airports (ORD, JFK,EWR, LGA, ATL, PHL) (Figure 6-10).

In summary, there is one cluster (Cluster 2) where neither the airport or link delay state variables play an important role; consequently, the time-of-day variable is the most significant variable and the classification error is high. There are two clusters for links that originate at one of the highly congested airports (Clusters 1, and 3). In one of these clusters, only the departure airport departure delay plays an important role (Cluster 1), while for the other one, both the departure airport's departure delay, and the prediction link's departure delay are important (Cluster 3). Finally, we have two clusters (Clusters 5 and 5) whose member links are destined for large, highly congested airport. In one of them (Cluster 4), the arrival airport's arrival delay and the prediction link's departure delay are the main variables, and in the other (Cluster 5), only the prediction link departure delay plays an important role.

We also investigate whether the location of a link has an impact on the clustering. If we look closely at Figure 6-10 we see that only links in the Eastern US belong to Clusters 4

Table 6.2: Clusters' centroid variables values for five clusters.

Clustering Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
	(24 elements)	(16 elements)	(21 elements)	(7 elements)	(32 elements)
Test error	18.6	20.6	19	19.9	18.6
Time-of-day	72.8	100	81.2	74.7	71.1
Day-of-week	5.6	6	6.3	7.2	6.2
Season	2.8	3	3.2	3.7	3.1
NAS state	17.2	19.9	25	10	18.3
NAS type-of-day	21.8	22.1	25.9	33.5	37.4
NAS prev. type-of-day	16.7	16.9	19.1	21.9	17.9
Dep. airport's dep. delay	94.8	30.4	79.7	18.6	17.7
Arr. airport's arr. delay	0.9	14.4	4.3	87.4	28.1
Pred. OD pair's dep. delay	41.3	53.3	82.3	86.8	95.5
Pred. OD pair's arr. delay	0	4.2	8.7	16.5	16.1
Highest airport's importance	98.4	44.4	85.3	87.4	38
2^{nd} highest airport's importance	67.6	33.9	56.5	72.2	32.3
3^{rd} Highest airport's importance	32.5	29.9	39.4	37.2	26.6
Highest OD pair's importance	56.3	62.7	95.3	86.8	100
2^{nd} Highest OD pair's importance	46.4	42.8	63.5	67.3	71.6
3^{rd} Highest OD pair's importance	41	36.5	48.2	53.2	55.1

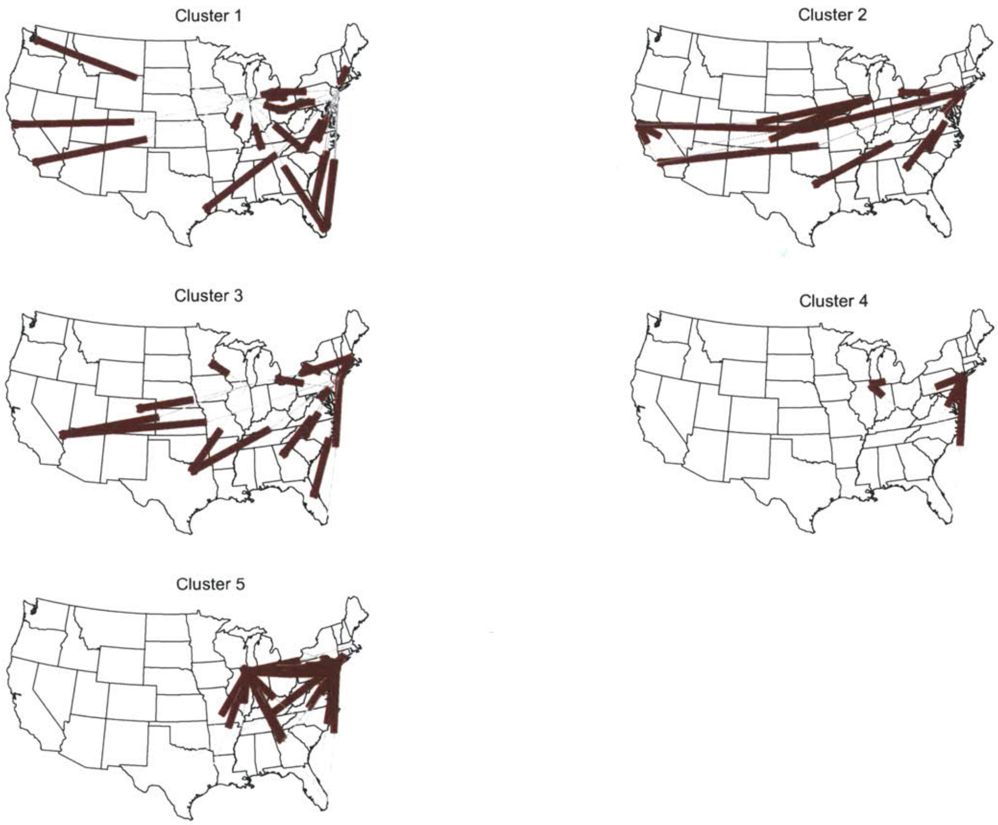


Figure 6-10: Clusters 1-5 OD pairs geographical location.

and 5. Arrivals to major delay centers which are not located in Eastern US typically belong to Cluster 2, where none of the airports' or links' delay state variables were found to be significant. On the other hand, if we compare Clusters 1 and 3 (departures from high delay centers), we see that links having ORD as origin belong to Cluster 1 when their destination is on one of the coasts, and those links with ORD as origin and destined for the South West belong to Cluster 3.

Finally, we applied the same clustering algorithms to the regression problem as well. The results for five clusters and the same clustering variables were almost identical, and the qualitative description of the centroids were very similar.

6.3 Identification of the most influential OD pairs

In this section, 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.

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

Origin	Destination	Total Importance
EWR	ORD	1713
STL	ORD	1553
MCO	EWR	1455
ATL	EWR	1363
ORD	EWR	1289
ORD	LGA	1218
LGA	ORD	1169
FLL	EWR	972
JFK	BOS	703
IAH	ORD	700

We now select three of the most influential links in Table 6.3, and take a closer look at the links for which they play an important prediction role. In Figures 6-11, 6-12 and 6-13, we see the links for which EWR-ORD, STL-ORD and FLL-EWR 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. The first aspect we want to highlight are the differences in the locations of the links for which these three links play an important role. For the EWR-ORD link, Figure 6-11 shows that the origins and destinations are not limited to a specific area of the US. This suggests that the EWR-ORD delay level is a good reflection of the global network delay state, since the level of delay of this link plays an important role in predicting the future delay of links with very different locations. Figures 6-12 and 6-13 show that the locations of the links for which STL-ORD and FLL-EWR delay variables play important prediction roles. We find that these links are more localized, suggesting that FLL-EWR and STL-ORD are good descriptors of the local delay situation.

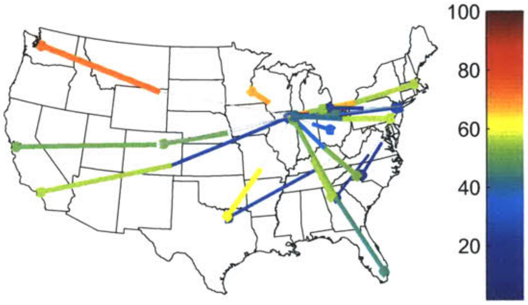


Figure 6-11: EWR-ORD explanatory variable importance for the 100 most-delayed OD pairs.

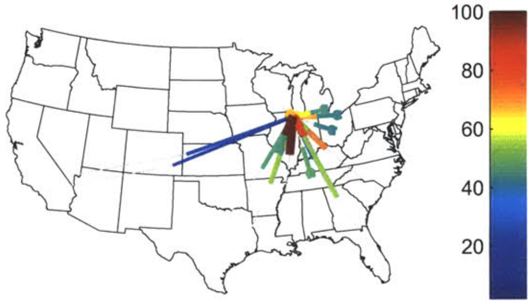


Figure 6-12: STL-ORD explanatory variable importance for the 100 most-delayed OD pairs.

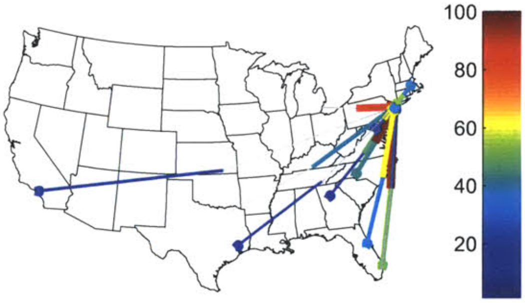


Figure 6-13: FLL-EWR explanatory variable importance for the 100 most-delayed OD pairs.

It is reasonable to find that the most influential links play an important predictive role in other proximate links' prediction models, since factors such as convective weather would affect them in similar ways. However, in some cases we have links' delay variables playing important roles for links that are not in their vicinity, but with a common departure or arrival airport. For example, the EWR-ORD departure delay state has a 72.4 importance value in the ORD-SEA departure delay prediction model. One possible explanation for this type of behavior is the aircraft routing, which can create dependencies between different links' delays. In Figure 6-14, we have the aircraft rotation information for aircraft flying from ORD to SEA. The arrows in the map depict where aircraft departing from ORD for SEA, flew into ORD from (2007-2008 data). Most of the aircraft arrived from SEA, (1,720). However, there are a significant number of aircraft coming from the East Coast; for example, 531 from LGA and 382 from BOS. The EWR-ORD delay state is a good indicator of the delay that these aircraft flying from the East Coast to ORD suffer. Aircraft rotations and link locations can help us understand why some links delay state variables play an important role in other links' prediction models. there are, however, instances with no obvious explanation for the appearance of a certain link delay state as a good prediction variable. There are network effects and correlations that the RF output help us to identify, but are difficult to explain. For example, the FLL-EWR delay plays an important role in the DTW-EWR delay prediction model, with an importance level of 81.9. The two links are not in the same geographic area, and the aircraft rotations information does not show strong aircraft connectivity between FLL and DTW: 4,379 of the aircraft flying from DTW to EWR came from EWR, and only a few aircraft came from the FLL area (Figure 6-15). We interpret the high importance value of the FLL-EWR delay variable in the DTW-EWR departure delay prediction model as a sign of the ability of the FLL-EWR delay variable to reflect high delay situations in the South East of the US, which are likely to affect the North East area (including DTW-EWR) two hours in the future.

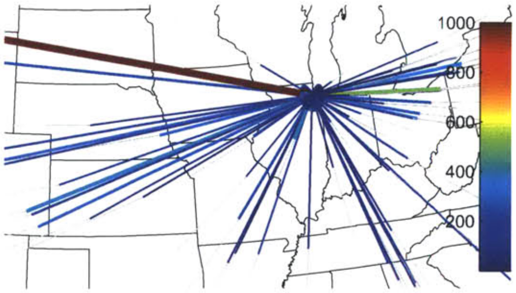


Figure 6-14: Number of flights on link preceding ORD-SEA in aircraft rotations (2007-2008 data).

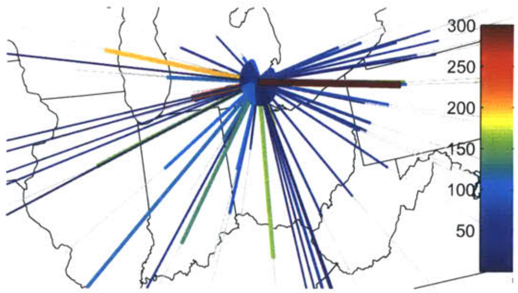


Figure 6-15: Number of flights on link preceding DTW-EWR in aircraft rotations (2007-2008 data).

6.4 Comparison of the OD pairs with the best and worst prediction performance

Previously, we saw that there are significant differences between the prediction models' performance for different OD pairs. For example, the classification test error (2h horizon, 60 min threshold) of the EWR-ATL departure delay prediction model is 11.3% (the lowest test error we obtained), while that of the LAS-SFO link test error is 28.8% (the highest test error). The median test errors for regression are also significantly different for these two links, 15.6 min, and 24.3 min respectively. The goal of this section is to compare these two models, and understand what makes the prediction performance so different. We focus on the role of the most significant prediction variables (importance level over 50).

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 6-16 and 6-17. They show the EWR-ATL and LAS-SFO 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 EWR-ATL confidence intervals overlap less with the 60 minute threshold line than the LAS-SFO intervals. The more the overlap and lower the distance from the intervals' center to the 60 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 LAS-SFO confidence intervals in Figure 6-17 are wider than the EWR-ATL intervals. This indicates lower correlation between the departure delay and the time-of-day variable, and an increased overlap with the threshold line.

both models had only one more variable with a significant importance level. This was the ATL-EWR departure delay variable for the EWR-ATL model (importance 87.3), and the LAS-SFO departure delay variable for the LAS-SFO model (importance 71.9). We calculated the correlation between the output variable (future delay) and the value of this variable, and the results were 0.67 for the EWR-ATL model, and 0.43 for the LAS-SFO model. Figure 6-18 and 6-19 show two 2D histograms, illustrating that the ATL-EWR variable is a better predictor of the future delay since it presents a higher correlation with

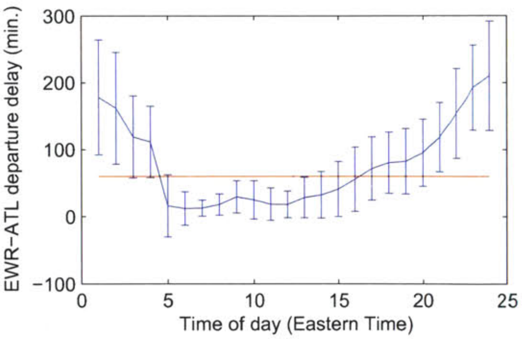


Figure 6-16: EWR-ATL mean delay by time-of-day ($\pm\sigma$).

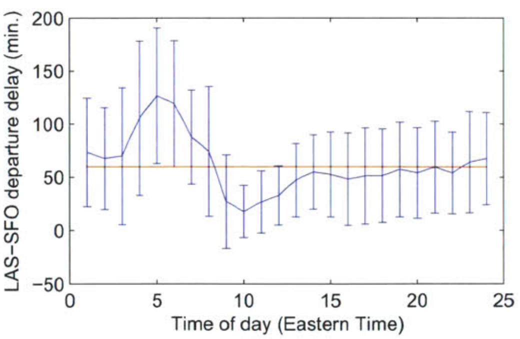


Figure 6-17: LAS-SFO mean delay by time-of-day ($\pm\sigma$).

it.

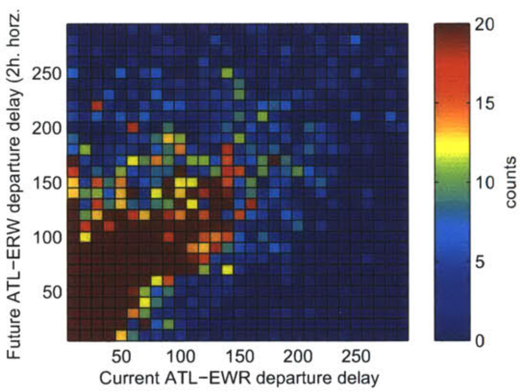


Figure 6-18: 2D histogram of EWR-ATL future departure delay versus current EWR-ATL departure delay.

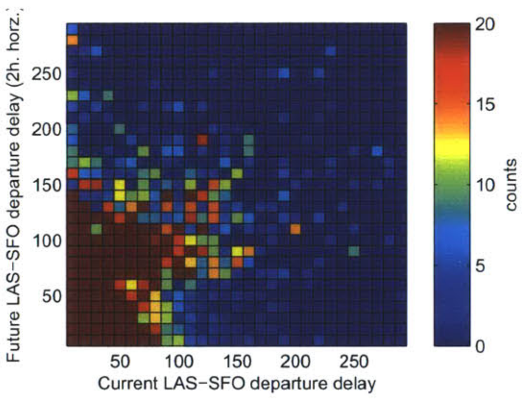


Figure 6-19: 2D histogram of LAS-SFO future departure delay versus current LAS-SFO departure delay.

The high variability of the departure delay for specific values of the time-of-day variable, and the lack of correlation between any of the airports' or links' delay state variables and the output variable (future delay) lead to the poor performance of the LAS-SFO prediction model. We believe that aircraft rotations are an important factor behind the low correlation values, since 63% of the aircraft flying from EWR to ATL flew previously from ATL to EWR, but only 14% of the aircraft flying from LAS-SFO flew previously from SFO to LAS (SFO the most frequent previous departure airport). There is no a strong correlation between the prediction models' performance and the percentage of flights arriving at the departure airport from the most frequent previous origin (termed the aircraft connectivity aggregation level). Figure 6-20 shows the test error versus the aircraft connectivity aggregation level for the 100 most-delays links. For example, we consider the ORD-LGA model. We see that the ORD-LGA test error and aircraft connectivity aggregation levels are both low. In the ORD-LGA prediction model, the ORD-LGA departure delay variable is the most important variable, while none of the others play an important role. The aircraft connectivity aggregation level in the ORD-LGA model is similar to that of LAS-SFO, but the prediction performance is low for LAS-SFO prediction. The low aircraft connectivity aggregation level does no appear to affect the prediction power of the ORD-LGA departure variable. This could be a consequence of the high impact of the ORD-related factors departure delays, and the long turnaround times of flights departing from ORD, which prevents delays from propagating.

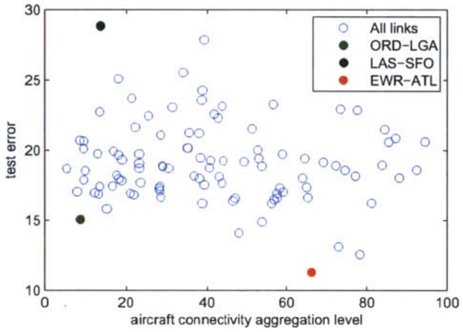


Figure 6-20: Test error versus aircraft connectivity aggregation level for the 100 most-delayed OD pairs.

6.5 Network aggregated error analysis

In this section, we treat the 100 most delayed links as one unique link. The test data for the 100 links is considered as a single test set. This leads to a 100x(10x1,000)=1,000,000 data points in the aggregated test set (10 test sets of 1,000 points each for each link). The goal is to identify general trends in the prediction models, to better understand the behavior of the models' prediction errors. This analysis would also enable us to compare individual link errors with the general trends. The results presented below were obtained for a classification problem with 60 min threshold and 2 hour horizon. We focus on the analysis of variations of the test error as a function of different variables.

Figure 6-21 shows the test error as a function of the future departure delay value. We see that the error is maximum around the classification threshold, when the error becomes close to 0.5 (random guess). The error rate decreases faster when we move from the maximum to the left, than when we move to the right; for example for 60-40=20min we have a 12% error, and for 60+40=100min we have 20% error. It is worth noting that the error rate increases for high values of the future delay (300min). Previously in the JFK-ORD prediction model, we attributed this effect to the lack of data points; however, Figure 6-21 once again suggests that the error rate increases for high delay values. This behavior shows that the delays of the different elements in the network have difficulties in explaining extremely high delays. The reason is likely these very high delays are associated with flights that had some kind of mechanical issue, or that were rescheduled later in the day due to a high delay situation, and by the time they departed, network delays were no longer high.

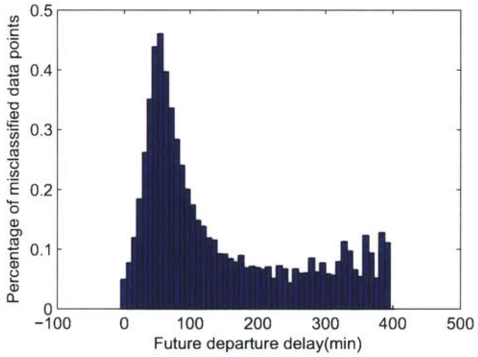


Figure 6-21: Test error versus future departure delay (network aggregated).

Figures 6-22 and 6-23 show the variation of the test error with the time of day, and the FPR/FNR ratio information respectively. Very early in the day (Eastern Time) the test error is low, being 5.7% at 4am. It and it increases steadily until 1pm, when a 23% error rate is reached. After 1pm the error rate decreases, but at 11pm it increases again. This leads to two local maxima in the test error versus time of day plot. Figure 6-22 also depicts the average delay for all the links in the simplified network versus time of day, and we can see that the two test error maximums take place when the average delay on the network is around 15 minutes. This is a transition value located in between the highest and lowest delay values. The error rate maximums occur when delays are harder to predict for our model, which is either when delays start increasing, and there is no clear sign of high delay in the network (the first maximum), or when delays start decreasing and there are still signs of high delay in the network (second maximum). There is also a clear trend in the FPR/FNR values versus the time of day, namely, as the day progresses the ratio increases. The FPR/FPR ratio is highly correlated with the average delay curve in Figures 6-22, and the correlation coefficient between these two quantities is 0.86.

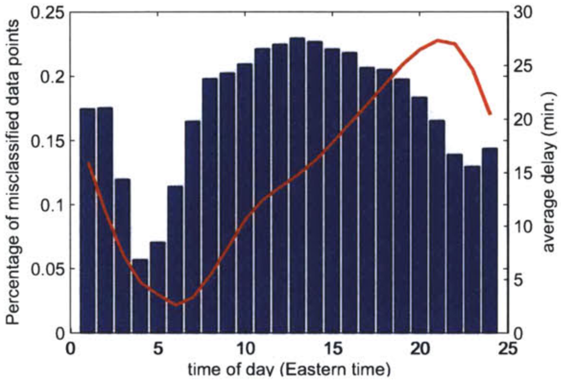


Figure 6-22: Network aggregated misclassification rate versus time of day

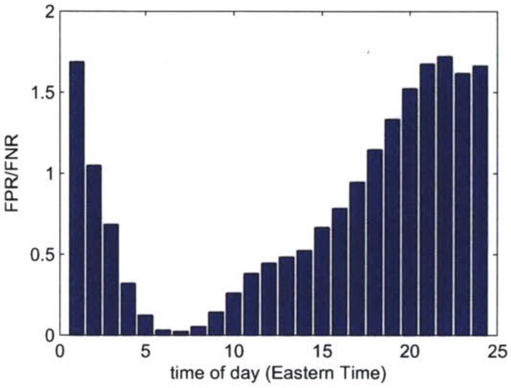


Figure 6-23: Network aggregated FPR/FNR versus time of day

The JFK-ORD test error versus time of day plot that we saw in the previous chapter (Figure 5-13) was a similar to the one presented here. However, the second maximum in Figure 5-13 was clearly dominant over the first one, suggesting that it is specially difficult to determine when delays start decreasing in the JFK-ORD link.

Figure 6-24 shows the average test error values for different months of the year showing their effect on the prediction. The minimum test error is reached in October (17.4%), and the maximum in February (21.3%). Higher test errors tend to occur in the summer and winter months.

6.6 Effect of changes in the classification threshold

This section considers the impact of changes in the classification threshold on the performance of the prediction models. We test three classification thresholds: 45, 60, and 90 min. The prediction time horizon is maintained at 2 hours.

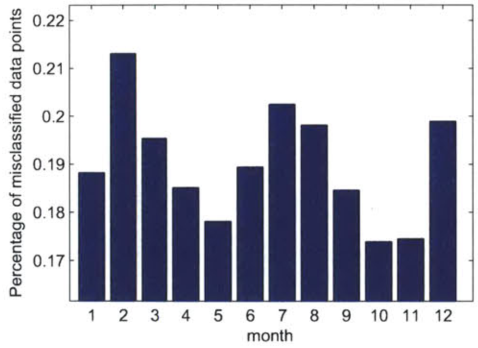


Figure 6-24: Network aggregated test error versus month of the year.

For the 100 most delayed links and the 45 min threshold, we obtain a mean test error of 21.2%, for the 60 minute threshold, the misclassification test error is 19.1%, and for the 90 minute threshold, 16.38%. The test error decreases as the classification threshold increases, since there are clearer indications of whether the future delay will exceed 90 min, than thee are for whether it will exceed 45 min.

Next, we look in detail at the values of the test error for the 100 most-delayed links. Figure 6-25 depicts the test error values for the three 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 6-26 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 2 pp to 8 pp. Table 6.4 shows the test error details for the three links with the largest and smallest error increase when moving from a 90 min. threshold to a 45 min threshold. The PHL-MCO link presents the largest error increase (8.7pp), and MIA-ORD the smallest error increase (2.1pp).

The importance of the explanatory variables is also affected by changes in the classification threshold. The major change takes place on the time-of-day explanatory variable. The average value of the time-of-day variable's importance for the 45 min threshold is 86.8, for the 60 min threshold is 78.5, and for the 90 min threshold is 67.7. The larger the threshold, the lower the importance value of this variable. As we mentioned in the previous

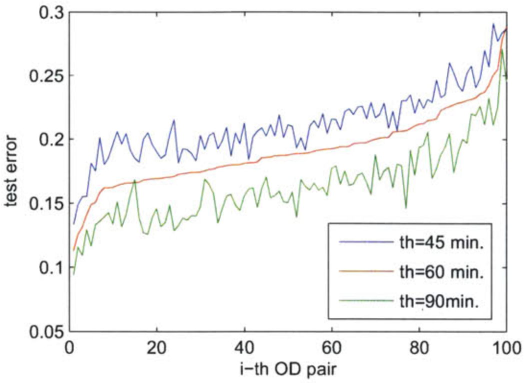


Figure 6-25: Classification threshold analysis.

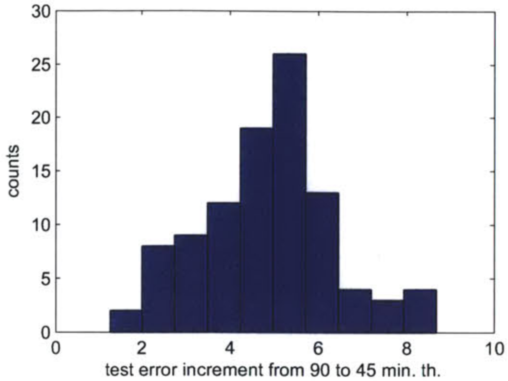


Figure 6-26: Histogram of the test error increment when changing the classification threshold from 90 min to 45 min.

Table 6.4: Test error changes with classification threshold.

OD pair	Test error (th=90min)	Test error (th=45min)	Error Increase
PHL-MCO	12.8%	21.5%	8.7
PHL-RDU	14.9%	23%	8.4
PHL-CTL	21.1%	29%	7.9
JFK-SFO	16.9%	19.2%	2.3
ORD-EWR	16.1%	18.4%	2.3
MIA-ORD	21.9%	24%	2.1

chapter, the time-of-day variable has difficulties explaining high delay values, which do not occur very often. However, for medium or low delay values which are exceeded quite often at a certain time of day, the time-of-day variable has a high predictive power. We did not identify any significant changes in the importance values of other explanatory variables.

6.7 Effect of changes in the prediction horizon

One would expect that the length of the prediction horizon would affect prediction performance. We measure the impact of the prediction horizon length on the classification and regression problems. We analyze four different time horizons: 2, 4, 6 and 24 hours. The classification threshold is maintained at 60 min.

The average classification test errors for the 100 most-delayed links at different time horizons are the following: 19.1% (2h), 21.4% (4h), 22.6% (6h), and 27.2% (24h). The average test error increase from 2 to 6 hours is only 3.5 percentage points. If we calculate the average test error for a model in which the only explanatory variable is the time-of-day, we obtain an average test error of 30%. The difference between this test error and the 24-hour horizon model test error is mostly due to the predictive value of the previous day's delay information. Figure 6-27 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.

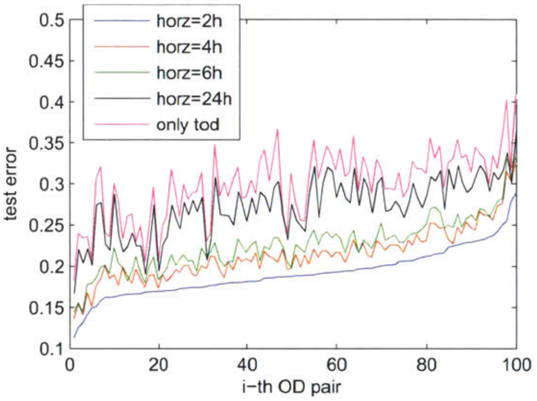


Figure 6-27: Classification prediction horizon analysis.

Table 6.5 summarizes the most interesting changes in the explanatory variables' importance with the time horizon. We can see that the time-of-day increases in importance with the time horizon. This is a consequence of the loss of prediction power of the airport, and link delay state variables as the time horizon increases. The NAS type-of-day's importance increases with the time horizon; however, for the 24 hour horizon this value drops, since we assume that we know the current type-of-day with certainty before the day is finished. This assumption makes the NAS type-of-day's importance higher as the prediction time horizon increases. On the other hand, for a 24 hour horizon the role of the type-of-day variable changes, it now depicts the previous day, and not the current day. For this reason the importance value decreases significantly for the 24 prediction horizon.

Finally, we present the results of the regression problem. The average median test error for the 100 links and the different time horizons are the following: 20 min (2h), 23 min (4h), 24.3 min (6h), and 27.4 min (24h). In other words, the average median test error

Table 6.5: Effect of changes in the prediction horizon on the explanatory variables average importance value (100 most-delayed OD pairs).

Explanatory Variable	**2h horizon**	**4h horizon**	**6h horizon**	**24h horizon**
Time of day	78.5	95.9	98.1	100
NAS type of day	28.5	52.6	60.8	17.4
Top 3 airports avg. importance	49.3	44	35.9	37.2
Top 3 links avg. importance	62.6	56	38.6	24.7

increase from 2 to 6 hours is only 4.3 minutes, and only 7.4 min as the prediction horizon increases from 2 to 24 hours. Figure 6-28 presents the median test error for all links and prediction horizons studied, as we presented in Figure 6-27 for classification. With respect to the regression variables' importance, the same trends identified in Table 6.5 apply to the regression problem.

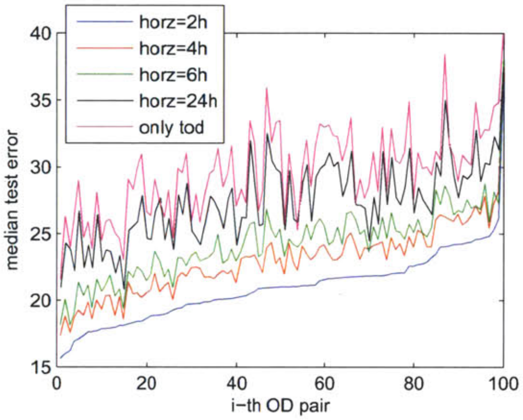


Figure 6-28: Regression prediction horizon analysis.

6.8 Departure delay prediction for scheduled flights

So far in this thesis, we predicted the departure delay of flights that actually take off at a time $t+T$, being T the prediction horizon length. In this section, we evaluate the performance of our model in predicting the departure delay of flights *scheduled* to depart at time $t+T$ instead of flights actually departing at time $t+T$.

With respect to the JFK-ORD departure delay model which we developed in Chapter 5, the classification test error for scheduled times a 2h-horizon, and a 60min classification threshold is 24.76%. The corresponding test error for actual departure times is 21.2%. The prediction error for a 2h-horizon with scheduled times is close to the 6h-horizon error for actual times, namely, 25.12%. Looking at the median regression error, we see that the 2h-horizon performance for scheduled time is also comparable to the 6h-horizon error for actual departure times (29.5).

For the 100 most-delayed links, we see a similar test performance decrease on average. For example, for actual times the classification test error increases by 3.5pp between a 2h and a 6h-horizon, and by 4.5pp between actual and scheduled departure times for a fixed 2h-horizon. However, as we can see in Figure 6-29, different links behave differently. This figure shows the classification performance loss when the prediction horizon increases from 2 to 6 hours for actual times versus the performance loss of scheduled times against actual departure times for a fixed 2h-horizon. We see that not all links are located near the 45 degree line. For example, for the JFK-LAX link (denoted by a red dot in the figure), the change in the time horizon increases the test error by 7.9pp, and the change from actual to schedule times (for a fixed 2h horizon) only increases the test error by 1.4pp.

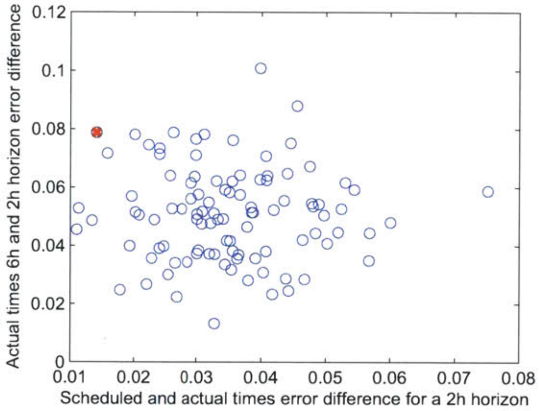


Figure 6-29: Comparison of the impact on the classification test error of changing the prediction time horizon from 2 to 6 hours versus using actual or scheduled departure times for a fixed 2h-horizon.

We have seen that using scheduled departure times instead of actual times increases the prediction error. By using scheduled times, we are predicting the delay of flights departing later which is similar to increasing the time horizon: The scheduled departure time plus the associated delay is the actual departure time. If we compare the importance values of the explanatory variables for scheduled and actual times for a fixed 2h-horizon, we once again see that using scheduled times is similar to increasing the time horizon. For example, the time of day variable average importance value is 78.5 for actual times, and 88.5 for scheduled times. As we saw previously in Section 6.7, an increase of the time horizon translates to an increase in the importance of the time of day variable.

Chapter 7

Conclusions & Next Steps

This thesis presented a new network-based air traffic delay prediction model that incorporated both temporal and network delay states as explanatory variables. The first step was to identify a simplified network which only contained links with significant traffic (more than 10 flights per day on average). ORD and ATL appeared as the airports with the most high-traffic links in the simplified network, with 90 and 82 links respectively. With the purpose of obtaining a good estimate of the delay state of each of the links and airports in the network, individual flight data were aggregated using a moving median filter with a 2h window and a 1h time step.

The prediction models presented in this thesis included temporal variables (time of day, day of week, season) and network delay state variables. We differentiated between two different types of network variables, namely, local delay state and high-level network delay state variables. The local delay state variables or influential airports and links were obtained using a RF based algorithm. These variables allowed us to identify several interesting interactions. For example, DCA departure delay showed up as the airport delay variable with the highest prediction power in the JFK-ORD departure delay prediction model. The high-level delay variables were obtained through clustering of the delay values of the simplified network links. The resultant NAS delay state clusters depicted ORD, NYC (EWR, JFK and LGA) and ATL airports as the main delay sources in the NAS. The NYC high delay cluster presented the highest average link delay among all clusters: 42.2 minutes. The main delay sources that showed up in the type-of-day clusters were the same as the ones that were identified in the NAS delay state analysis. Both the NAS state and the type-of-day variable temporal analysis showed that ORD high delay state takes place more often in the winter months, and that the NYC and ATL high delay states are more frequent in the summer months. We also saw that September, October, and November are the months when high delay states occurs less frequently, and when the low delay state is more frequent.

In this thesis we predicted departure delays. We are interested in comparing different links prediction performance, and this is hard to do for arrival delays due to the dependence of the prediction horizon with the length of the link. For example, if we want to predict the arrival delay of a flight two hours before departure time, the length of the prediction time horizon will be 2 hours plus the travel time, and the travel time depends on the link length.

Of prediction models evaluated in this thesis, the RF algorithm showed the best performance, and was selected in our study. We tested the RF classification and regression delay prediction model on the 100 most-delayed OD pairs in the NAS. The goal was to study the performance of prediction models for different OD pairs, and to identify sets with similar characteristics (for example, similar prediction performance or explanatory variables importance). The results obtained showed an average test error of 19% when classifying delays as above or below 60 min, at a 2-hour prediction horizon, and a 20.9 min median test error for regression. The analysis also found that the dependence of individual link delays on the network state varied from link to link. The 100 most-delayed OD pairs were clustered in 5 groups of links sharing similar explanatory variables' importance value and performance. This allowed us to identify some interesting correlations, for example, that an increase of the time-of-day variable's importance is associated with an decrease of the prediction performance. We also identified the OD pairs that had the most influence on the 100 most-delayed OD pairs EWR-ORD was found to be the most important link, followed closely by STL-ORD.

The results presented in this thesis quantified the effects of the classification threshold and the prediction horizon on the predictive performance of the models. Both the classification and regression models were found to be quite robust to increases in the prediction horizon: The median regression test error (averaged across the 100 most-delayed OD pairs) only increased from 20 min to 27.4 min when the prediction horizon increased from 2 hours to 24 hours, and the classification test error increased from 19.1% to 27.2% for the same change in the prediction horizon. For a fixed 2h prediction horizon for the 100 most-delayed links and a 45 min classification a mean test error of 21.2% was obtained. For a 60 min threshold, the misclassification test error was 19.1%, and for a 90 min threshold, 16.4%.

The NAS delay state variables proposed in this paper enabled the development of the promising network-based delay prediction models. These variables could potentially be used in the development of a network delay prediction and analysis tool. For example, we could use the previous day's type-of-day information to help us predict the current day's NAS delays. Other next steps in this research include studying the effect of changes in the moving median filter parameters. We could test different window sizes and step sizes, and evaluate the impact of those changes on the prediction models' performance. New explanatory variables could also be added to the prediction models. By following a strategy similar to the one we used to obtain the NAS delay state categorical variable, we could introduce a new categorical variable depicting the delay state, not of the entire network, but of a certain area in the vicinity of the link of interest. This variable would allow us to eliminate local delay state variables from the prediction models while possibly achieving the same performance.

Bibliography

[1] Bureau of Transportation Statistics, http://www.transtats.bts.gov/HomeDrillChart.asp, 2012.

[2] Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio Trani, Bo Zou, *Total Delay Impact Study*, 2010.

[3] Joint Economic Committee (JEC), *Your Flight has Been Delayed Again: Flight Delays Cost Passengers, Airlines, and the US Economy Billions*, 2008.

[4] S. AhmadBeygi, A. Cohn, Y. Guan and P. Belobaba, *Analysis of the potential for delay propagation in passenger airline networks*, Journal of Air Transport Management, Vol. 14, Issue 5, September 2008, pp. 221-236.

[5] M. Jetzki, *The propagation of air transport delays in Europe*, Thesis, Department of Airport and Air Transportation Research, Aachen University, 2009.

[6] Y. Tu, M. O. Ball and W. S. Jank, *Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern*, American Statistical Association Journal, 2008, vol. 103, pp 112-125.

[7] R. Yao, W. Jiandong and X. Tao, *A flight delay prediction model with consideration of cross-flight plan awaiting resources*, ICACC, 2010.

[8] B. Sridhar and N. Chen, *Short term national airspace system delay prediction*, Journal of Guidance, Control, and Dynamics, Vol. 32 No. 2, 2009.

[9] Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., and MacPhail, T., *Predicting Weather Impact on Air Traffic*. ICNS Conference, Herndon, VA, May 2007.

[10] N. Xu, K. B. Laskey, G. Donohue 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.

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

[12] J. Rice, *Mathematical Statistics and data analysis*, 3rd ed., Duxbury Press. 2006.

[13] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, 2nd ed., Springer 2009.

[14] G. Upton and I. Cook, *A Dictionary of Statistics*, 2nd ed., Oxford University Press 2008.

[15] Q. Xu and Y. Liang, *Monte Carlo cross validation*, Chemometrics and Intelligent Laboratory Systems, Vol. 56, Issue: 1, Pages: 1-11, 2001.

[16] C. Fan, H. Wen, W. Yirong, *The Comparison of the V-Fold and the Monte-Carlo cross validation to estimate the number of clusters for the fully polarimetric sar data segmentation*, Geoscience and Remote Sensing Symposium, 2007.

[17] J. FurnKranzs, E. Hullermeier, *Preference Learning*, Springer 1998.