Transportation Research Part C 44 (2014) 231–241

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Characterization and prediction of air traffic delays

Juan Jose Rebollo, Hamsa Balakrishnan *

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 5 June 2013
Received in revised form 16 April 2014
Accepted 16 April 2014

Keywords:
Air traffic delay prediction
Network effects
k-Means clustering
Random Forests
Classification
Regression

ABSTRACT

This paper presents a new class of models for predicting air traffic delays. The proposed models consider both temporal and spatial (that is, network) delay states as explanatory variables, and use Random Forest algorithms to predict departure delays 2–24 h in the future. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and links (origin–destination pairs) in the network, new network delay variables that characterize the global delay state of the entire National Airspace System at the time of prediction are proposed. The paper analyzes the performance of the proposed prediction models in both classifying delays as above or below a certain threshold, as well as predicting delay values. The models are trained and validated on operational data from 2007 and 2008, and are evaluated using the 100 most-delayed links in the system. The results show that for a 2-h forecast horizon, the average test error over these 100 links is 19% when classifying delays as above or below 60 min. Similarly, the average over these 100 links of the median test error is found to be 21 min when predicting departure delays for a 2-h forecast horizon. The effects of changes in the classification threshold and forecast horizon on prediction performance are studied.

1. Introduction

Flight delays in the United States result in significant costs to airlines, passengers and society. The annual cost of domestic flight delays to the US economy was estimated to be $31–40 billion in 2007 (Ball et al., 2010; Joint Economic Committee, US Senate, 2008). Such high delay costs motivate the analysis and prediction of air traffic delays, and the development of better delay management mechanisms (Manley and Sherry, 2010; Ferguson et al., 2013; Glover and Ball, 2013; Delgado et al., 2013).

The large number of shared resources in the National Airspace System (NAS) together with aircraft, crew and passenger interdependencies result in the propagation of delays through the network (AhmadBeygi et al., 2008; Jetzki, 2009). The desire to maximize aircraft utilization reduces the time buffer between arrivals and departures, increasing the likelihood of delay propagation (AhmadBeygi et al., 2008). Increasing demand also decreases the ability of the network to absorb disruptions, making the network susceptible to large-scale delays. The study of network effects can help identify factors that mitigate or amplify delay propagation in the NAS.

Delay prediction has been the topic of several previous efforts. Jetzki (2009) studied the propagation of delays in Europe, with the goal of identifying the main delay sources. Tu et al. (2008) developed a model for estimating flight departure delay distributions, and used the estimated delay information in a strategic departure delay prediction model. Yao et al. (2010)

* Corresponding author. Tel.: +1 617 253 6101.
E-mail address: hamsa@mit.edu (H. Balakrishnan).

http://dx.doi.org/10.1016/j.trc.2014.04.007
0968-090X/

focused exclusively on downstream delays caused by aircraft, cockpit and cabin crew connectivities. By contrast, Bratu and Barnhart (2005) focused on the impact of delays on passengers. Recently, Pyrgiotis et al. (2013) have considered delay propagation in a network of airports using a queuing model. Other prediction models (Klein et al., 2007; Klein et al., 2010; Sridhar and Chen, 2009) have focused on weather-related delays, and the development of a Weather Impacted Traffic Index (WITI). Xu et al. (2005) proposed a Bayesian network approach to estimating delay propagation. Using a system-level Bayesian network, the authors were able to capture interactions among airports. None of these prior approaches have investigated the role of a network delay state in predicting future delays. By contrast, the goal of this paper is to evaluate the potential of network-scale delay dependencies in developing delay prediction models.

Due to the existence of network effects, current delays in the NAS are expected to be a good indicator of the short term evolution of delays in the system. It is therefore useful to determine state variables that reflect the current situation, and use them to predict future delays. The models presented in this paper therefore attempt to predict future departure delays on a particular origin–destination (OD) pair by considering current and/or past delays in the network. The proposed prediction models will not capture delays that only affect a few aircraft (for example, mechanical delays). The objective of these models is *not* to predict individual flights delays, but instead to estimate the future network-related delay on a certain route. However, it is important to note that the models are evaluated using actual data containing all delays, including those that impacted only a few flights. The prediction models presented in this paper yield a better understanding of delay interactions between the different elements in the NAS. They also help assess how much of the future delay on a particular link can be explained by the current delay state of the network.

2. Problem definition

The main objective of this paper is to predict the departure delay on a particular link or at a particular airport, some time in the future. The departure delay of a link at time t is an estimate of the departure delay of any flight(s) taking off at time t, and flying on that link. For example, if the BOS-MCO departure delay state two hours from now is estimated to be 30 min, it means that the estimated departure delay for BOS-MCO flights taking off two hours from now is 30 min. Two types of prediction mechanisms are considered: *classification*, where the output is a binary prediction of whether the departure delay is more or less than a predefined threshold, and *regression*, where the continuous output is an estimate of the departure delay along the link.

2.1. Data sources

The results presented in this paper were obtained using data from the Aviation System Performance Metrics (ASPM) database (Federal Aviation Administration, 2012), for the two-year period beginning January 2007 and ending December 2008. The ASPM database provides detailed data for individual flights by phase of flight, airport weather data, runway configuration, and arrival and departure rates. The fields used in the analyses in this paper include the arrival and departure airports, the scheduled and actual gate-in times, the scheduled and actual wheels-off times, the flight carrier codes, and the aircraft tail number.

The individual flight data were processed to obtain a more robust aggregate delay estimate. A moving median filter, a low-pass filter, was used to obtain the delay states of airports and OD pairs. The delay state of a 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. The window size, W, was set to two hours, and t was incremented in steps of one hour.

The raw data set spanned 2029 airports and 31,905 OD pairs, most of which averaged fewer than one flight a day. Since the analysis focused on network effects, only OD pairs with at least 10 flights per day on average were considered. Fig. 1 shows the resulting simplified network, which is composed of 112 airports and 584 OD pairs.

3. Characterization of the network delay state

The interdependencies among the different elements in the NAS and repetitive traffic patterns support the development of characteristic NAS delay states, that reflect the current situation at a network-level. These states are characterized both as a "snapshot" in time (i.e., the current delay patterns in the NAS), and in terms of temporal evolution (i.e., a characteristic type of day). In addition to yielding insights into system behavior, these delay states can be used as explanatory variables in prediction models.

3.1. Characteristic NAS delay states

The observed NAS departure delay at time t is a 584-dimensional vector, defined by the departure delay state of each link in the simplified network at time t. The 17,519 NAS departure delay observations were classified into a few typical NAS delay states using the k-means clustering algorithm (Hastie et al., 2009).

The k-means algorithm partitions the observations in the data set into k clusters so as to minimize the sum of distances within each cluster. In the case of departure delays, the centroids of each of the clusters represent the typical NAS delay

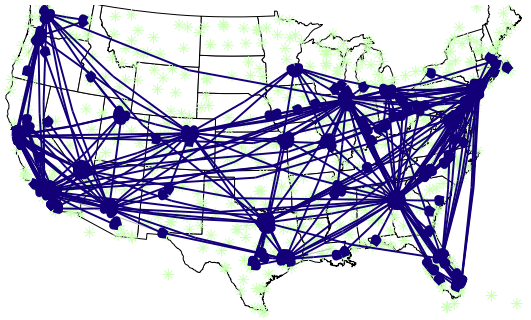

Fig. 1. Simplified network for the continental US. The green icons denote airports in the original dataset that are not part of the simplified network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

states. Six clusters were chosen because, as seen in Fig. 2, the total intra-cluster distance did not decrease much for more than 6 clusters. Figs. 3 shows the cluster centroids' delay levels, illustrating that Chicago (ORD), New York City (NYC) and Atlanta (ATL) are the main delay centers. For fewer than six clusters, ATL does not appear; for more than six clusters, no new delay centers appear, further supporting the choice of $k = 6$. Table 1 qualitatively describes the six identified delay states, along with the average link delay of its centroid and the number of observations classified as belonging to it.

3.2. Characteristic types of delay days

Similar to the identification of characteristic delay states, typical types of delay days can be identified by clustering an entire day's worth of time series data using the k-means algorithm. In this case, each observation contains $584 \times 24 = 14{,}016$ variables (i.e., number of OD pairs $\times$ 24 h/day). Fig. 4 shows the total distances and the within-cluster distances for different numbers of clusters.

The main delay centers were found to be the same as seen in the NAS delay state clusters: ORD, ATL, and NYC. Table 2 describes the main source of delay at the highest delay point of the day for each of the six identified type-of-day clusters, along with the average delay in the NAS for that type of day, and the number of observations belonging to it.

Fig. 5 shows the monthly occurrences of each type of day in 2007–2008. Day 1 (high NYC delays, and significant ORD and ATL delays) was more common in the summer months, while Day 4 (high NYC delays, but not high ORD or ATL delays) was seen year-round, with higher frequency around the summer months. The ORD high delay day (Day 2) was found to be more frequent in winter, while the ATL high delay day (Day 3) was more frequently seen in summer.

4. Delay prediction

Ten training sets (3000 points each) and ten test sets (1000 points each) were sampled from the 2007–2008 data set. The prediction models were fit and tested for each of the 10 training and test set pairs, respectively, providing measures of the variability and test error. The training and test sets were over-sampled from the 2007–2008 data. Over-sampling is the over-selection of samples of the minority class in order to achieve balanced training and test datasets with sufficient representatives of both the majority and minority classes (Upton and Cook, 2008). Since the majority of links do not experience delays

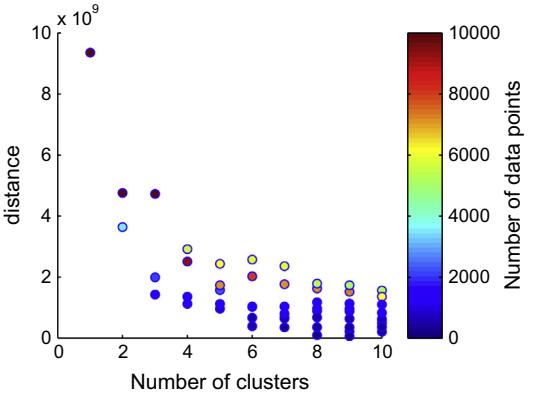

Fig. 2. (Left) Intra-cluster distances vs. number of clusters.

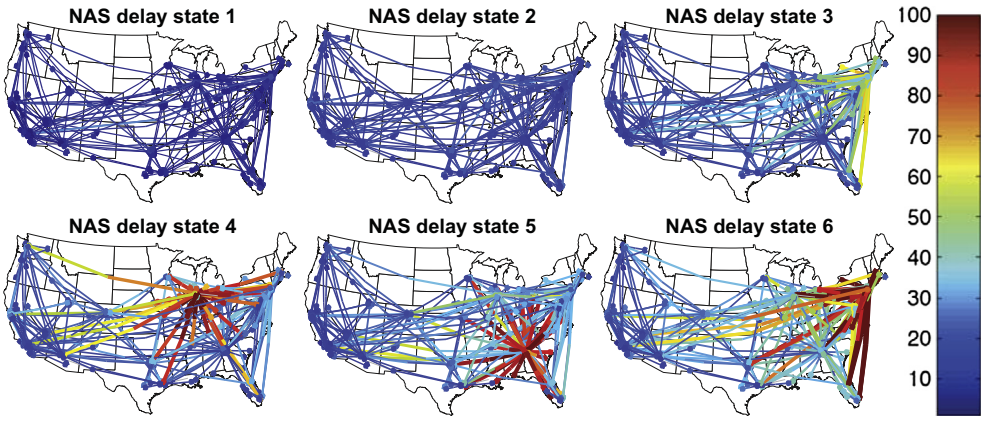

Fig. 3. Centroids of NAS delay states for six clusters.

Table 1
NAS delay state clustering. Delay definitions: high (90 min), medium–high (60 min), medium (20 min), low (5 min).

Delay state	Qualitative description	Avg. centroid link delay (min)	Elements in cluster
1	NAS low delay	5.8	8029
2	NAS medium delay	15.2	5915
3	NAS medium high delay	24.4	1505
4	ORD high delay	31.2	1192
5	ATL high delay	32.9	398
6	NYC high; ATL, ORD medium delay	42.2	480

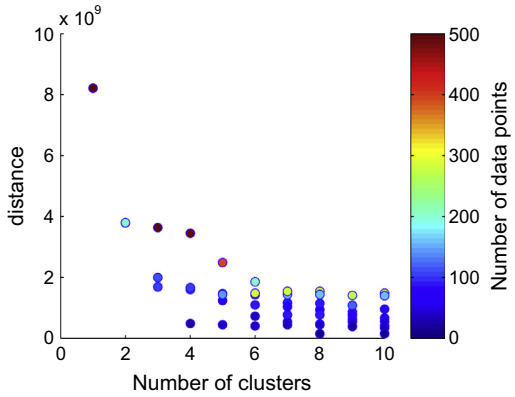

Fig. 4. Intra-cluster distances vs. number of clusters.

Table 2
Descriptions of the types of NAS delay days. Delay definitions: high (90 min), medium–high (60 min), medium (20 min), low (5 min).

Day type	Qualitative description	Avg. centroid link delay (min)	Elements in cluster
1	NYC high+; ATL, ORD high delay	29	31
2	ORD high; NYC medium–high delay	22	94
3	ATL high; NYC, ORD medium–high delay	21	29
4	NYC high; ATL, ORD medium delay	19	86
5	NYC, ORD medium delay	15	207
6	NAS low delay	9	282

of more than 60 min, a naïve classification algorithm that predicts no delays of more than an hour would be correct most of the time. For this reason, the true evaluation of a classifier's performance is its ability to correctly predict delays in a balanced data set in which half the points have delays of less than 60 min (the so-called "majority class"), and half the points have delays of more than 60 min (the "minority class").

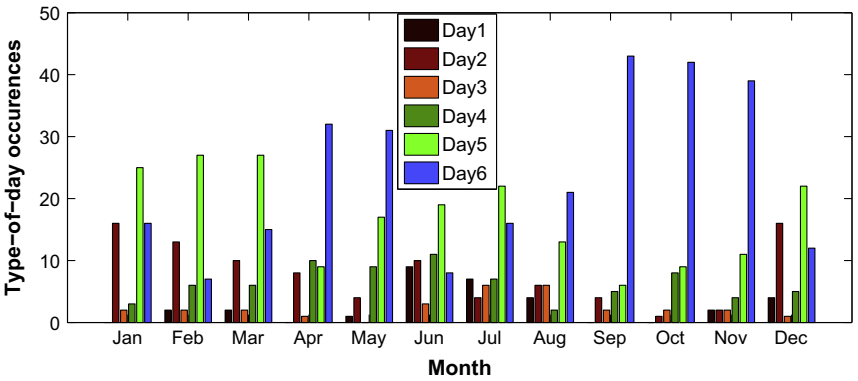

Fig. 5. Occurrences of NAS type-of-day by month in the 2007–2008 dataset.

Different classification and regression models (logistic regression, single classification trees, bagging, boosting, linear regression, neural nets and Random Forests) were tested, and Random Forests were chosen due to their superior performance (Hastie et al., 2009; Rebollo, 2012).

4.1. Random Forests

The Random Forest (RF) is an ensemble classifier that consists of many decision trees, and outputs the mode of the classes output by individual trees (Breiman, 2001). It combines the concept of bagging with the random selection of variables at each tree split (Hastie et al., 2009). Benefits of Random Forests include the automatic generation of variable importance, their low sensitivity to outliers in the training data, and their good performance in cases where the number the variables is large compared to the number of samples. The RF approach has also been extended to regression (Breiman, 2001).

5. Explanatory variables

The delay prediction models proposed in this paper use a combination of categorical and continuous explanatory variables. The Kruskal–Wallis parametric ANOVA test (Rice, 2006) and the multiple comparisons test were used to evaluate the dependence of the future departure delay with different categories for the proposed categorical explanatory variables. A parametric ANOVA test was used because of the highly-skewed nature of the delay distributions. The Random Forest methodology was used to identify the most relevant continuous variables.

The level of significance of the explanatory variables is expected to vary depending on the desired output of the prediction model (regression vs. classification), as well as the forecast horizon. It is, however, reasonable to assume that if an explanatory variable is significant for the regression problem, it will also be significant in the classification problem.

5.1. Temporal variables

Temporal variables considered included the time-of-day, the day-of-week and the month-of-year, all of which were treated as categorical (the time-of-day variable has one category for each hour). The low p-values obtained by the ANOVA tests, as well as the multiple comparisons tests, showed that delays varied significantly with all three temporal variables. Based on the analysis, the month-of-year was replaced by a "season" variable with three categories: September–November (with typically experience low delay levels), January–May (moderate delay levels), and June–August and December (high delay levels).

5.2. Spatial variables

5.2.1. NAS delay state, type of delay day, and previous day's type

The primary spatial variables considered were the NAS delay state and the type of delay day. In practice, delay information will only be known from the beginning of the day until the time when the prediction is made. However, in the analyses in this paper, the type of day is assumed to be known with certainty at the time of prediction, and its importance is evaluated. NAS delays for the previous day are known with certainty, and can help predict delays on a given day. The NAS does not immediately recover from high delay situations, such as, a day with strong convective weather and 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. The multiple

Table 3
Influential airports and OD pair delay variables for JFK-ORD departure delay prediction. The importance values are normalized such that the most influential variable of each kind has an importance of 100.

Apt. delay variable	Importance	OD pair delay variable	Importance
DCA departure	100	JFK-ORD departure	100
JFK departure	96.9	EWR-ORD departure	90.9
ORD arrival	85.3	LGA-ORD departure	65.3
ORD departure	82.8	ORD-JFK departure	44
LGA departure	58.9	ORD-LGA departure	24.3
BOS departure	58.9	BOS-ORD departure	17
PHL departure	58.2	PHL-ORD departure	16.9
EWR departure	57.7	JFK-FLL departure	11.9
JFK departure	56.3	BUF-JFK arrival	11.4
DCA arrival	46.1	LGA-ORD arrival	11

comparisons test results showed significant differences for the variable reflecting the previous day's type; this variable was therefore included in the prediction model.

5.2.2. Delays at influential airports and OD pairs

Future delays on a particular OD pair are likely to be influenced by delays at certain airports or OD pairs. For example, while predicting departure delays on the JFK-ORD link, it is reasonable to expect that ORD departure delays and JFK arrival delays will play an important role. The model considers arrival and departure delays at 400 airports, which results in 800 variables, which are ordered according to their ability to predict delays on a particular link. The importance of different variables, as determined by the RF algorithm, was used to choose the most relevant airport delay variables. A similar analysis was conducted for the 1064 OD pair delay variables. Table 3 shows the 10 airport delay and OD pair delay variables used in the JFK-ORD departure delay prediction model. The variable importance is normalized by the most important variable of that type. Influential delay variables are similarly identified for each OD pair in the simplified NAS network.

Not surprisingly, Table 3 suggests that most of the important variables reflect the delays prevalent in the NYC and ORD areas. There are two potential explanations for the high importance of the DCA departure delay variable. Firstly, DCA is located close to the route between JFK and ORD. Airspace weather events that have a large impact on the JFK-ORD route will also affect DCA airport delays. The routes served by the airports as another important element. DCA serves more short haul, East Coast flights (similar to JFK-ORD) than either JFK or ORD, which serve International traffic. The average DCA delay therefore reflects the local delay state, and consequently the JFK-ORD delay state. The latter hypothesis is supported by the absence of IAD (which is close to DCA but serves international traffic) in the list of influential airports for JFK-ORD. Another interesting finding is that the JFK-FLL departure delay has nearly the same importance as the BUF-JFK and LGA-ORD arrival delays when predicting the JFK-ORD departure delay. It is conjectured that factors such as aircraft routings and airline networks drive these features.

6. Departure delay prediction for the 100 most-delayed OD pairs

The performance of the prediction models are evaluated for the 100 OD pairs in the simplified NAS network with the highest average delays. Classification and regression predictors were developed for each OD pair. The performance of the classification-based and regression-based departure delay prediction models were first studied for a 2-h prediction window and a 60 min classification threshold (that is, predictions of whether the delay will be above or below 60 min).

6.1. Classifier performance

The performance of the classification models is first considered. Fig. 6 shows the test error histogram for the 100 most delayed OD pairs. The test error ranges from 11.3% to 28.8%, and the average value is 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%). 90% of the analyzed links have a test error standard deviation under 1.7 percentage points. Fig. 6 (right) shows the empirical cumulative distribution function (cdf) of the test error standard deviations of all 100 links.

Further inspection of the test error reveals that the False Negative Rate (FNR) clearly dominates the False Positive Rate (FPR), that is, the classifier is more likely to misclassify a high delay link than to predict high delay when in reality the delay on the OD pair is low. For the 100 most-delayed links, the average FNR is 23.62% and the average FPR is 14.6%; the FNR rate is higher than the FPR for all OD pairs. This behavior is because the proposed prediction model focuses on the delay state of the different elements in the network, but does not capture localized delays (for example, mechanical issues). As a result, even if the delays on the relevant network elements were low, an OD pair may experience high delays 2 h later due to a local issue that only affects a certain flight, but this delay would not be predicted by the proposed model.

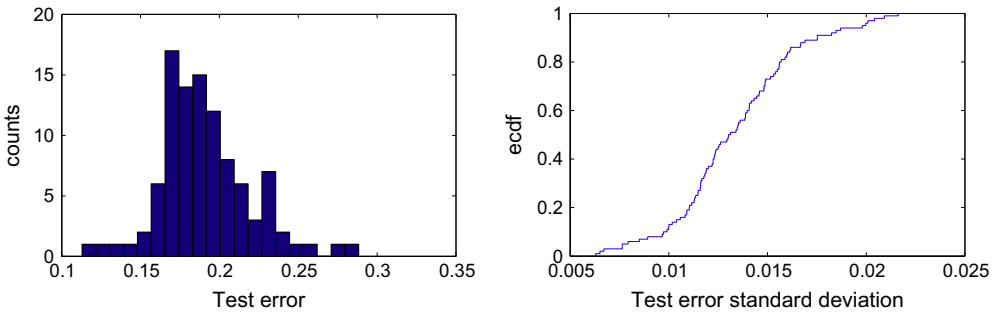

Fig. 6. (Left) Classification test error histogram and (right) empirical cumulative distribution function (ecdf) of the standard deviation of the classification test error, for the 100 most-delayed OD pairs.

Fig. 7 shows the FPR and FNR versus the test error for all the studied OD pairs. For the lowest test error OD pair, FNR/FPR ratio is 1.3, while for the highest test error OD pair (LAS-SFO), FNR/FPR = 1.9, showing that the FNR dominance increases with the test error.

Time-of-day is the most important explanatory variable for both the OD pair with the lowest (EWR-ATL) and the highest (LAS-SFO) test errors. The large difference in the test errors can be explained using Fig. 8. The figure shows the EWR-ATL (left) and LAS-SFO (right) departure delay means and one standard deviation intervals versus the time-of-day, for the test data. The EWR-ATL confidence intervals overlap less with the 60 min threshold line than the LAS-SFO intervals. The more the overlap and the less the distance from the mean values 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. The LAS-SFO confidence intervals in Fig. 8 are wider than the EWR-ATL intervals, indicating less correlation between the departure delay and the time-of-day variable. Delays for flights arriving or departing from SFO are hard to predict: The average test error rate for links that have SFO as origin or destination is 23.3%. This behavior is likely a result of reduced airport capacity due to fog at SFO, and the resulting delays that can linger though the day.

6.2. Regression performance

The performance of the regression approach is evaluated using the same data set that was used in Section 6.1. Fig. 9 (left) shows the histogram of the median test error for the 100 most-delayed OD pairs. 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 can be seen in Fig. 9 (right), the standard deviation of these error values is low, with the 90th percentile of the error distribution being 1.17 min. A significant gap is seen between the highest median error value (LAX-HNL), and the second highest (SFO-JFK). Since neither of these links had the highest test error in classification, a natural question is whether links with high classification test error also have high regression test error. Fig. 10 addresses this question by showing the classification error versus the regression error. 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 Fig. 10 corresponds to the CLT-LGA departure delay prediction model. The classification test error in this link is high (22.6%, and in the 87th percentile of the classification error distribution), but the regression median test error is only 20.2 min (in the 40th percentile of the regression error distribution). In general, the classification and regression problems are different: The former requires information that helps to

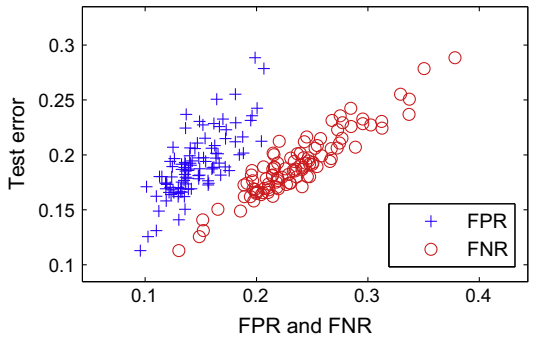

Fig. 7. FNR and FPR scatter plots for the 100 most-delayed OD pairs.

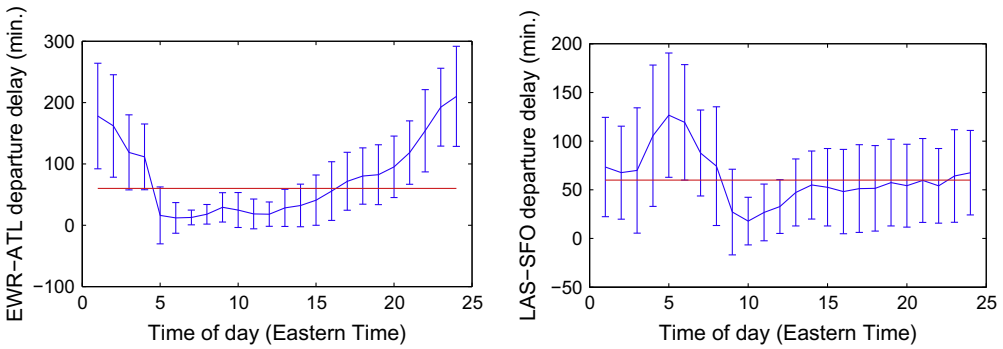

Fig. 8. (Left) EWR-ATL and (right) LAS-SFO mean delays by time-of-day ($\pm\sigma$).

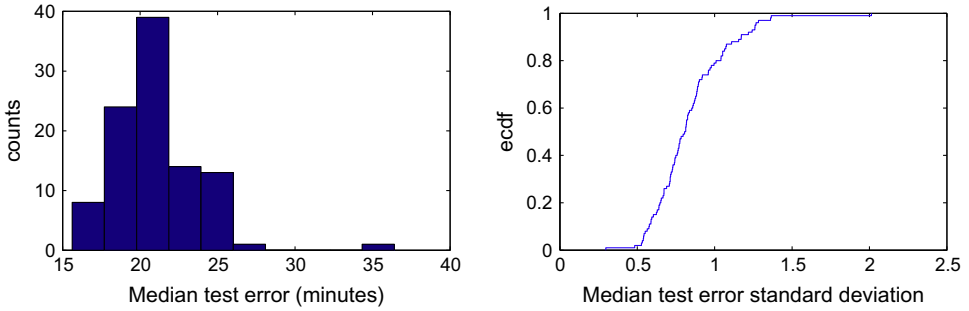

Fig. 9. (Left) Regression median test error histogram. (Right) Empirical cdf of the standard deviation of the regression median test error.

differentiate between high and low delays, whereas the latter tries to predict the *value* of the future delay. Depending on the classification threshold and the typical delays on a link, the relative performance on classification and regression can differ.

6.3. Effect of changing classification threshold

Three values of the classification threshold, namely, 45, 60, and 90 min, are studied for the 100 most-delayed links, and a forecast horizon of 2 h. For a 45 min threshold, the mean test error is found to be 21.2%; for the 60 min threshold, the misclassification test error is 19.1%; and for the 90 min threshold, 16.4%. As expected, the test error decreases as the classification threshold increases.

Fig. 11 (left) shows the test error values for the three threshold values for each of the 100 most-delayed links; the links are ordered according to their test error for the 60 min threshold. This plot shows that not all links have the same reduction in test error when increasing the classification threshold, and that the reduction is not correlated with the value of the test error. Fig. 11 (right) depicts the histogram of the increase in test error when moving from a 90 min threshold to a 45 min threshold in classifying delays. The increase in test error ranges from 2 percentage points to 8 percentage points, depending on the OD pair.

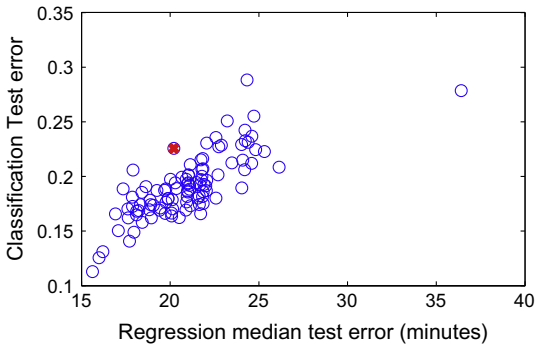

Fig. 10. Classification vs. regression test error.

6.4. Effect of increasing forecast horizon

The prediction performance would be expected to decrease with an increase in the forecast horizon. The performance on the classification and regression problems is analyzed for forecast horizons of 2, 4, 6 and 24 h. The classification threshold is maintained at 60 min.

The average classification test errors over the 100 links and different forecast horizons are found to be 19.1% (for a 2-h forecast horizon), 21.4% (4-h forecast horizon), 22.6% (6-h forecast horizon), and 27.2% (24-h forecast horizon). By contrast, a model with only time-of-day as an explanatory variable yields an average test error of 30%. The difference between this test error and the 24-h forecast horizon test error is largely due to the predictive value of the previous day's delay information.

The regression problem shows similar results as the forecast horizon increases. The average median test errors over the 100 most-delayed OD pairs are found to be 20 min (for a 2-h forecast horizon), 23 min (4-h), 24.3 min (6-h), and 27.4 min (24-h). In other words, the average median test error increase is only 7.4 min as the forecast horizon increases from 2 to 24 h.

6.5. Prediction of delays relative to scheduled departure times

Thus far, the departure delay of an OD pair at time t was defined as the average delay of flights on that OD pair that *actually* take off at time t. In this section, the performance of the proposed model in predicting the departure delay of flights *scheduled* to depart at time t is evaluated.

For the JFK-ORD link, a 2-h forecast horizon and a 60 min classification threshold, the classification test error for predicting delays for scheduled departure times is 24.8%. For comparison, the corresponding test error for actual departure times was 21.2%. The prediction error for a 2-h forecast horizon with scheduled times is close to the 6-h forecast horizon error for actual times, namely, 25.1%. The median regression error for a 2-h forecast horizon with scheduled times is also comparable to that of the 6-h forecast horizon error for actual departure times (29.5 min).

On average, the 100 most-delayed links exhibit a similar deterioration in test performance when scheduled departure times are used instead of actual ones. For example, the classification test error increases by 3.5 percentage points between a 2-h and a 6-h forecast horizon for actual departure times, while it increases by 4.5 percentage points between the actual and scheduled departure times for a fixed 2-h forecast horizon. However, individual links behave differently. For example, the 2-h to 6-h increase in forecast horizon increases the test error of the JFK-LAX link by 7.9 percentage points, while the change from actual to scheduled times (for a fixed 2-h horizon) only increases the test error by 1.4 percentage points (Rebollo, 2012).

7. Implications of results

To the best of our knowledge, the proposed models present the first attempt to predict flight delays at a future time, using only aggregate variables that are presently available. For example, in the Bayesian network model proposed by Xu et al. (2005), the explanatory variables include the air carrier delay, Ground Delay Program holding times, departure demand, departure demand, airport capacity, en route weather report and air carrier rescheduling decisions, all corresponding to the flight for which delays are being estimated. However, these variables are typically not available prior to the actual flight time. Similarly, the Pyrgiotis et al. (2013) model tries to estimate the extent of air traffic delay propagation using a queuing network model consisting of the OEP-35 airports in the US. The authors acknowledge that the purpose of the model is not to reproduce the exact delays, but only the trends and behaviors that are seen (Pyrgiotis et al., 2013). Other models focus on

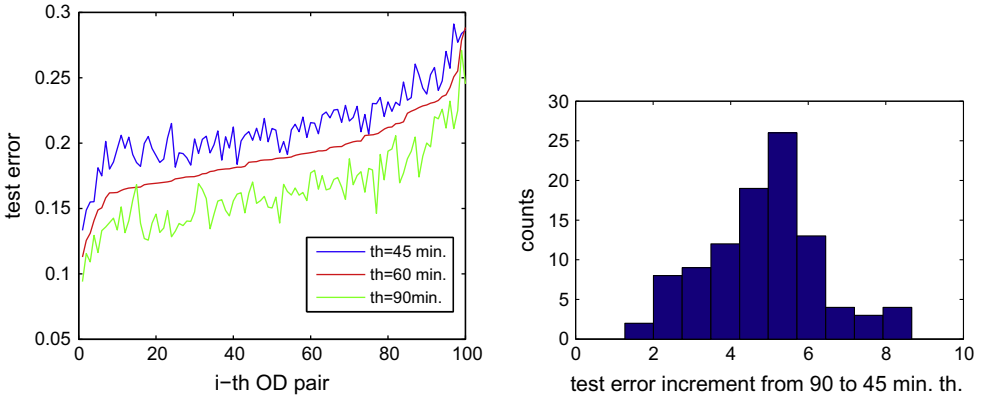

Fig. 11. (Left) Classification threshold analysis. (Right) Histogram of the test error increment when changing the classification threshold from 90 min to 45 min.

estimating, not predicting, weather impacts (Klein et al., 2007; Klein et al., 2010; Sridhar and Chen, 2009) or airline operational factors (Yao et al., 2010).

In evaluating classification algorithms, two commonly-used metrics are the accuracy and recall. In the present context, accuracy would be the percentage of correctly classified links, and recall would be the percentage of delayed links that are correctly classified as delayed. The aim of classification would be to maximize the recall, although this would result in a decrease in accuracy Hastie et al. (2009). From the results in Section 6.1, we note that our classification algorithm achieves an average accuracy of 81% and an average recall of 76.4%. These rates appear to be better than those reported anecdotally by commercial flight delay prediction tools (Skomoroch, 2009), although it is important to bear in mind that these tools predict individual flight delays on unbalanced datasets, while this paper focuses on predicting aggregate delays on balanced datasets.

Benchmarking of the regression performance is more difficult, due to a lack of suitable comparisons. However, we note that the US Department of Transportation (Bureau of Transportation Statistics, 2014) only counts a flight as delayed if it incurs a delay of more than 15 min, suggesting that a mean prediction error of 21 min may not be unreasonable.

The models developed in this paper serve two purposes. They help identify features that influence future delays, thereby improving our understanding of system behavior. In practice, accurately predicting delays 6 h or a day prior to the scheduled departure time could enable travelers make appropriate changes to their travel plans, such as changing their flight connections tactically, in order to improve their travel experience. For this reason, the next step in this research will be to see if the lessons learned in this research can be used to predict individual flight delays.

8. Conclusions

This paper presented new network-based air traffic delay prediction models that incorporated both temporal and network delay states as explanatory variables. The new models were enabled by the development of novel NAS delay state variables that represented the system as a whole. The results obtained for the 100 most-delayed OD pairs in the NAS showed an average test error of 19% when classifying delays as above or below 60 min, for a 2-h forecast horizon. The analysis also found that the dependence of individual link delays on the network state varied from link to link. Both the classification and regression models were found to be quite robust to increases in the forecast horizon: The median regression test error (averaged across the 100 OD pairs) only increased from 19.1 min to 27.4 min when the forecast horizon increased from 2 h to 24 h. Similar promising results were obtained for the prediction of departure delays relative to the scheduled flight times.

Acknowledgments

This work is supported in part by the NSF Cyber-Physical Systems projects ActionWebs (Award Number 0931843) and FORCES (Award Number 1239054).

References

AhmadBeygi, S., Cohn, A., Guan, Y., Belobaba, P., 2008. Analysis of the potential for delay propagation in passenger airline networks. J. Air Transport Manage. 14 (5), 221–236.
Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A., Zou, B., 2010. Total Delay Impact Study.
Bratu, S., Barnhart, C., 2005. An analysis of passenger delays using flight operations and passenger booking data. Air Traffic Control Quart. 13 (1), 1–27.
Breiman, L., 2001. Random Forests. Machine Learn. 45 (1), 5–32.
Bureau of Transportation Statistics, accessed 2014. Airline On-Time Performance Data, <transtats.bts.gov>.
Delgado, L., Prats, X., Sridhar, B., 2013. Cruise speed reduction for ground delay programs: a case study for San Francisco International Airport arrivals. Transport. Res. Part C: Emerg. Technol. 36, 83–96.
Federal Aviation Administration, 2012. Aviation System Performance Metrics database, <http://aspm.faa.gov/aspm/ASPMframe.asp>.
Ferguson, J., Kara, A.Q., Hoffman, K., Sherry, L., 2013. Estimating domestic US airline cost of delay based on European model. Transport. Res. Part C: Emerg. Technol. 33, 311–323.
Glover, C.N., Ball, M.O., 2013. Stochastic optimization models for ground delay program planning with equity-efficiency tradeoffs. Transport. Res. Part C: Emerg. Technol. 33, 196–202.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed.
Jetzki, M., 2009. The propagation of air transport delays in Europe. Master's thesis, Department of Airport and Air Transportation Research, Aachen University.
Joint Economic Committee, US Senate, 2008. Your Flight has Been Delayed Again: Flight Delays Cost Passengers, Airlines, and the US Economy Billions.
Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., MacPhail, T., 2007. Predicting weather impact on air traffic. In: Integrated Communication, Navigation and Surveillance (ICNS) Conference.
Klein, A., Craun, C., Lee, R.S., Airport delay prediction using weather-impacted traffic index (WITI) model. In: Digital Avionics Systems Conference (DASC).
Manley, B., Sherry, L., 2010. Analysis of performance and equity in ground delay programs. Transport. Res. Part C: Emerg. Technol. 18 (6), 910–920.
Pyrgiotis, N., Malone, K.M., Odoni, A., 2013. Modelling delay propagation within an airport network. Transport. Res. Part C: Emerg. Technol. 27, 60–75.
Rebollo, J.J., 2012. Characterization and Prediction of Air Traffic Delays. Master's thesis, Massachusetts Institute of Technology.
Rice, J., 2006. Mathematical Statistics and Data Analysis, third ed. Duxbury Press.
Skomoroch, P., 2009. How FlightCaster Squeezes Predictions from Flight Data. <www.datawrangling.com/how-flightcaster-squeezes-predictions-from-flight-data>.
Sridhar, B., Chen, N., 2009. Short term national airspace system delay prediction. Journal of Guidance, Control, and Dynamics 32 (2).
Tu, Y., Ball, M.O., Jank, W.S., 2008. Estimating flight departure delay distributions – a statistical approach with long-term trend and short-term pattern. Am. Stat. Assoc. J. 103, 112–125.
Upton, G., Cook, I., 2008. A Dictionary of Statistics, second ed. Oxford University Press.

Xu, N., Laskey, K.B., Donohue, G., Chen, C.H., 2005. Estimation of delay propagation in the national aviation system using bayesian networks. In: 6th USA/Europe Air Traffic Management Research and Development Seminar.
Yao, R., Jiandong, W., Tao, X., 2010. A flight delay prediction model with consideration of cross-flight plan awaiting resources. In: International Conference on Advanced Computer Control (ICACC).