Predictive Analytics for Airline Operations

Sohrab Mamdoohi [a], Hossein Fotouhi [a], Alexander R. Aguilar [a], Amber Jabeen [a], Yiyou Tan [a]
[a] *Volgenau School of Engineering, George Mason University, Fairfax, Virginia*

ARTICLE INFO

Article History:
Received 7 May 2020
Available 18 May 2020

Keywords:
Air transportation
Time series regression
Support vector machines
Random forests
Artificial neural network
Long short-term memory
Delay analytics

ABSTRACT

The National Airspace System (NAS) accommodates over 30 thousand flights per day. Each aircraft performs between 4 and 12 flights per day. The aircraft accumulate arrival delays as the day progresses due to waiting for other flights to depart from arrival gates, waiting for passengers and crew from connecting flights, waiting for aircraft services that are delayed servicing other flights, and a myriad of other factors related to departure and arrival slot availability. Dispatchers in Airline Operation Centers (AOC), and air traffic managers at the Air Traffic Control Command Center (ATCC), monitor the status of the cumulative flight arrival delays in the NAS. When arrival delays are anticipated, AOC/ATCC takes actions to mitigate the effects including cancelling flights, reprioritizing flights, and delaying departures. Their ability to manage the arrival flight delays is predicated by an accurate forecast at Noon of the cumulative arrival delays throughout the rest of the day.

This project analyzed the daily cumulative arrival flight delay data in the NAS for two years and evaluated four alternative methods for predicting cumulative arrival flight delay data for the rest of the day starting at Noon. The cumulative NAS arrival flight delays for 2018/2019 has mean = 293.81 ($\times 10^3$ minutes), median = 270.13, and standard deviation of 120.42. A cluster analysis identified three groups of cumulative NAS arrival flight delays: Normal (< 337.88), Semi-Normal (337.88 to 543.55), Abnormal (> 543.55). A Recursive Linear Regression method forecasting hourly cumulative delays performed with RMSE = 28.5 e03. Classification methods Support Vector Machine (SVM), Random Forest (RF), and Long Short-Term Memory neural networks (LSTM) method yielded accuracy of 0.772. 0.734, and 0.780, respectively. The details of the methods and limitations of each method are described in the report.

E-mail addresses:
smamdooh@gmu.edu (S. Mamdoohi), hfotouhi@gmu.edu (H. Fotouhi), aaguila8@gmu.edu (A. Aguilar), ajabeen@gmu.edu (A. Jabeen), ytan20@gmu.edu (Y. Tan).

1. Introduction

Every day individuals are traveling and taking flights to get to their destination in a timely manner. Airlines are in-charge of making sure that these flights are leaving at the appropriate time. Sometimes these flights can have a delay due to a disruption. These disruptions can be at the airport, in the airspace, with equipment issues/concerns, with crew members, and much more. A single disruption can have a ripple-effect on the entire network of flights. It is important to detect these disruptions, identify contributing factors, and quantify disruption-induced delay.

Airlines are required to manage sometimes thousands of flights a day. Travelers make travel plans, sometimes well in-advance, to their scheduled departure date. There is no knowledge of what potential disturbances can change a flight's departure and arrival time on that given day. Not only does one disruption affect a flight's departure and/or arrival, but that aircraft sequentially alters the departure and arrival time of flights afterward. Due to the delicate interdependency of required resources to operate a flight including but not limited to: aircraft, crew, and connection flights, disruption-induced delays may propagate in the airline network dramatically. With this in mind, a delay propagation modeling, developed through data analytics approach, is required in airline scheduling and management.

There have been many other researchers, statisticians, and others in the operations field to tackle airline flight delay prediction. There are different nuisances in how the researchers approached their analysis, whether they looked at a particular airport, or focused on a specific airline, the end goal remained the same, which was to provide flight delay predictions for the future with a palatable rate of accuracy. There are also many different statistical approaches taken with varying degrees of success. (Chakrabarty 2019) details a review of approaches taken in other studies to predict the arrival delay for American Airlines. Several methods and algorithms, such as the gradient boosting classifier model, were also used to achieve desired results. A validation accuracy of 85.73% was achieved by applying Grid Search on Gradient Boosting Classifier Model on flight data. (Khanmohammadi, Tutun and Kucuk 2016) takes a different route to predicting flight delays. The authors of this study are looking at predicting flight delays specifically at JFK airport by using different artificial neural network techniques. This study discusses the use of a multi-level input layer neural network, and specifically decides on the defect of module prediction – artificial neural network (DMP-ANN). During the course of the study, some limitations were found in this particular type of analysis. For example, there was not a way to assign any type of significance to say a

particular day was more important than another. The results of the analysis are that this method of DMP-ANN could be used for some problems related to airline flight delays, but not all. There was a lot of added complexity because of this model, so that would be addressed in future work.

After reviewing multiple approaches, one technique that possibly incorporates into the possible solutions is applying Long Short-Term Memory (LSTM) to predict airline flight delays. To the best of the authors' knowledge, LSTM has not applied for airline flight delay prediction purposes to date. The LSTM model will present some challenges, but also overcome some by solving some technical problems inherent to other statistical methods, such as vanishing gradients and exploding gradients (Phi 2018). They are particularly adept when solving problems that are too complex to solve using standard recurrent neural networks (RNNs). In (Greff, et al. 2017), the authors lay out their analysis on eight LSTM variants, in problems related to speech recognition, hand-writing recognition, and polyphonic music modeling, but recognizing that LSTMs are now applied to a variation in different types of problems. Throughout the study, the architecture of LSTMs is described and detailed, as well as its application to different problems. In the conclusion the authors described the simplification of several of the problems with the use of LSTM, but not as much in the performance and accuracy. This concluded the literature review giving the team ample resources to turn to.

Overall, the main objective of this study will be limited to identifying the end-of-the-day delay of the National Airspace System. Performance of well-known machine learning methods to predict the delay is tested. In addition to these methods, a novel approach for regression time series modeling is proposed and its performance is tested.

The rest of this study is organized as follows. In the second section, descriptions of the data, and methods applied to prepare the data are provided. The third section presents the modeling approaches in detail. The fourth section of this article reports prediction results and a comparison of the applied approaches. Discussion and conclusion about the performance of these approaches, and possible future works complete the article.

2. Data

Analyzing delays on a network-wide level is achieved through the use of a rich and precise dataset. The delay estimation model would be built on from the schedule and delay data. This section discusses the source of data. It also describes steps taken in the data preparation process and illustrates data visualizations for a better understanding of the dataset.

2.1. Data preparation

Historical airline on-time performance data were provided by the Bureau of Transportation Statistics (BTS) (Airline On-Time Performance Data 2020). The on-time performance of marketing carriers from January 2018 to December 2019 was obtained by the team for the study. This dataset provides 75 attributes with above 15 million observations. All the data corresponds to how the air transportation system is working throughout the network of airports within the US. Each record, which represents a flight, includes scheduled and actual departure/ arrival times. The total delay of each flight record is provided and further categorized into five classes of causes: 1. career, 2. weather, 3. national air system, 4. security, and 5. late aircraft. The dataset also contained other features such as flight origin and destination, airlines unique ID, flight number, tail number (the unique number associated with each aircraft), taxi-out/ in time, wheels-off/ on time, the distance between origin and destination, and canceled/ diverted flights. Since the dataset for each month was downloaded separately, the first step after downloading the dataset was to merge them. This step was performed in Python using one of Argo's computing nodes because of the large amount of memory needed.

In order to cleanse the dataset, the first step is to remove the outliers in the merged dataset. This is done by dropping the records associated with the flights which has the taxi-in and taxi-out times greater than 200 minutes, and the flights that have actual airtimes that are 100 minutes greater than their scheduled airtime.

All the timestamps provided in the dataset are given in local times. These timestamps include flights scheduled and actual arrival/departure dates and times. In the dataset, the scheduled arrival/departure times are reported as Computerized Reservations Systems (CRS) arrival/departure times. Also, each flight has an actual arrival and departure time. Based on the actual and CRS arrival and departure times, the arrival and departure delay of each flight can be calculated. This calculation, however, leads to a negative delay for some flights. Since we are interested in the positive delay, the negative delays are replaced with zero.

The next step is to convert all local timestamps to a reference time zone. In this project, all timestamps are converted to the Eastern Standard Time (EST). Converting the local times to EST is not trivial. However, to convert all timestamps, first the time zones associated with origin and destination cities of each flight are determined. This step cannot be done directly and is only feasible through obtaining the latitudes and longitudes of the cities. Afterward, time zones were identified by the coordinates. Next, given the time zone of each of the origin and destination of a flight, the timestamps are converted to EST. Since the flight arrival/departure times (both CRS and Actual) are not given in a standard format, a procedure is developed in Python to convert these times, which simultaneously changes the dates if it is necessary. This step was a crucial part of the project because the flight will be aggregated based on their actual arrival times, and it is important to calculate the flights' arrival and departure times precisely.

The final dataset contains the information of each flight including CRS arrival date, CRS arrival time, CRS departure date, CRS departure time, actual arrival date, actual arrival time, actual departure date, actual departure time, marketing airline network, origin airport ,origin city, destination, destination city, arrival delay, and departure delay.

2.2. Data exploration and visualizations

In this section, the dataset is explored visually by creating several plots and maps. Arrival delay is an effective factor in measuring air transportation network performance. It represents the efficiency of the air transportation system. Additionally, the first measure of performance that most passengers can realize is the arrival delay. The delay varies over time of the day, and is induced by types of causes such as career delay and late aircraft delay. We segmented each day into 24 one-hour time intervals and calculated the total arrival delay of the network for each one-hour time interval. To give an illustration of how delay varies over time of the day, the hourly and cumulative delay profiles of Jan 2019 was developed and shown in Figure 1. Additionally, the recurrent delay profile, which is the average delay under a normal condition (when there is no major incident to increase the delay substantially), was computed. For each time interval of a day (one-hour interval), the recurrent delay pattern was developed. To obtain the Recurrent Hourly Delay Pattern (RHDP) and Recurrent Cumulative Delay Pattern (RCDP), Loess Regression was applied. Unlike historical methods, which calculate the average delay for each time interval, Loess Regression is considered as unbiased by outliers. Both types of delay patterns are shown in red curves in Figure 1.

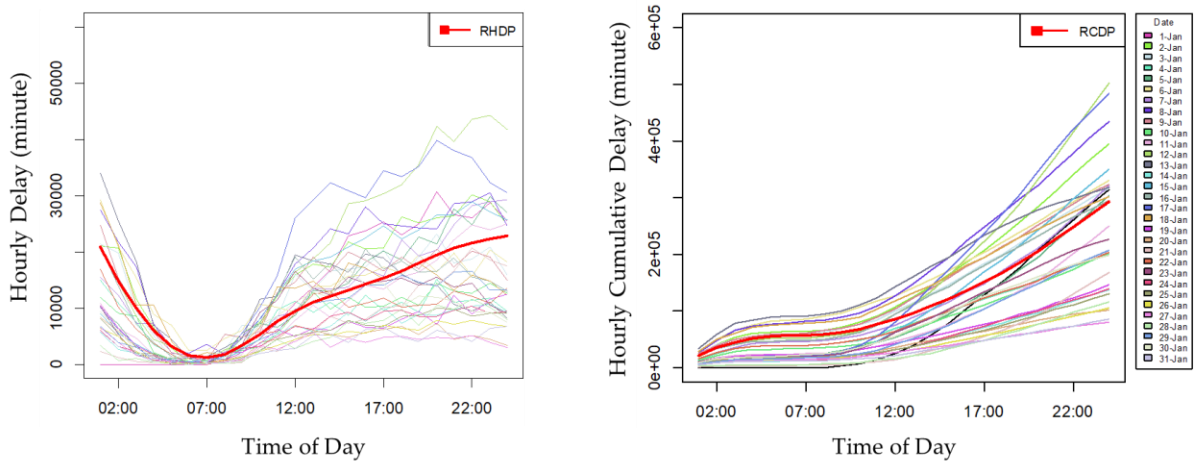


Figure 1. Profiles and recurrent patterns of: (a) hourly, and (b) cumulative delays.

We further developed the hourly delay profiles of the entire dataset (2018 and 2019), which is shown in Figure 2. It can be observed that hourly profiles of Jan 2019 are in line with those of the entire 2018 and 2019. In this figure, each line represents a day, and the delay of each hour is calculated by summing up the arrival delays of all flights which arrive during that hour of the day. As shown in this figure, the minimum hourly delay happens at 6 a.m., and the maximum delay happens at the end of the day (at midnight). Similarly, Figure 3 shows the cumulative hourly delay for all flights in 2018 and 2019.

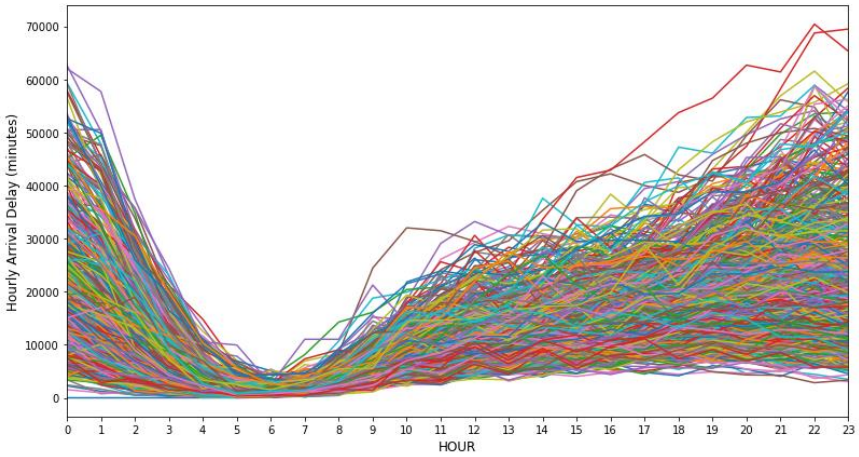


Figure 2. Arrival delay for all days in 2018 and 2019.

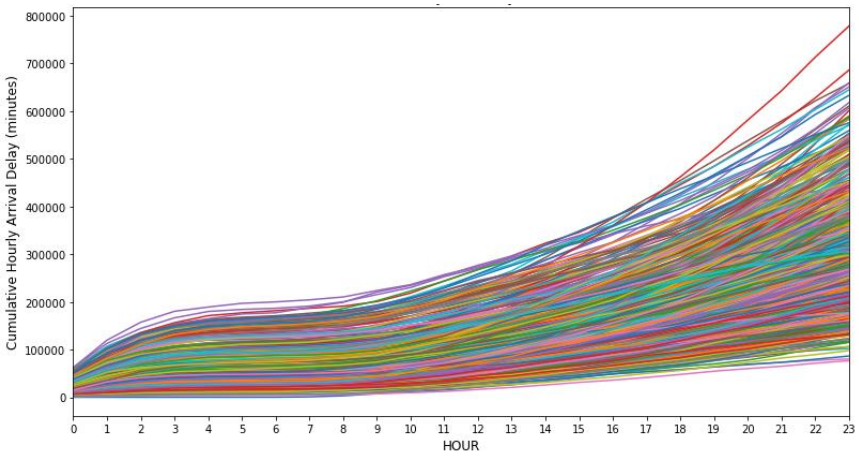


Figure 3. Cumulative arrival delay for all days in 2018 and 2019.

2.2.1. Delay Visualizations

In the following figures, maps were developed to visualize delays features. Figure 4 depicts the average delay of each flight over 2 years for selected airports. The highest average arrival delay belongs to Newark Liberty International Airport (EWR), with approximately 22 minutes. Chicago O'Hare International Airport (ORD) has the next highest average arrival delay with almost 20 minutes. Ted Stevens Anchorage International Airport (ANC) has the average arrival delay of one and a half (1.5) minutes, which is lowest among selected airports.

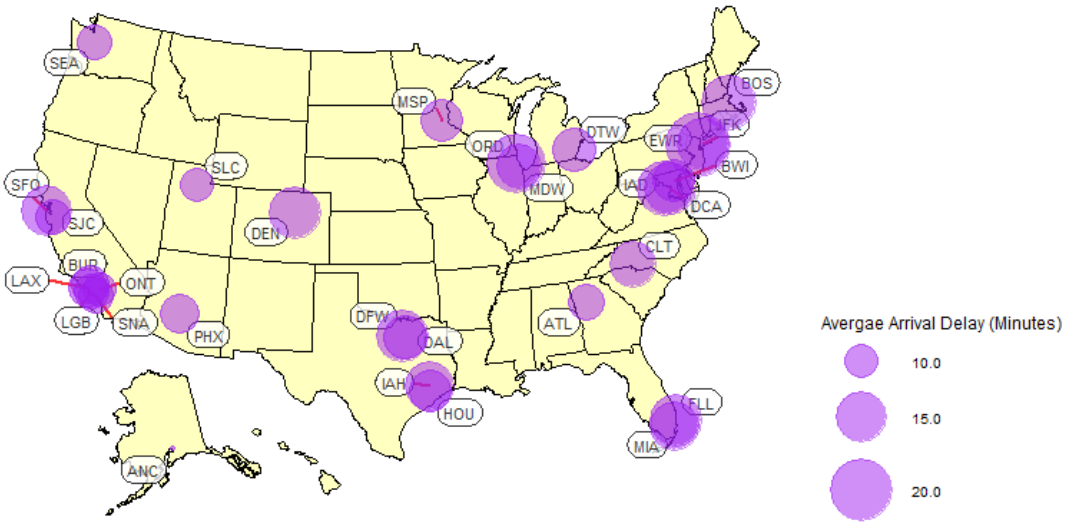


Figure 4. Average arrival delay of US airports.

To find the days and time intervals that are more subject to delay, a temporal delay heat map was created. As shown in Figure 5, it can be observed that flights arriving after 8:00 pm on Thursdays and Fridays are subject to higher delays. Flight routes between selected airports with respect to their average arrival delay are shown in Figure 6.

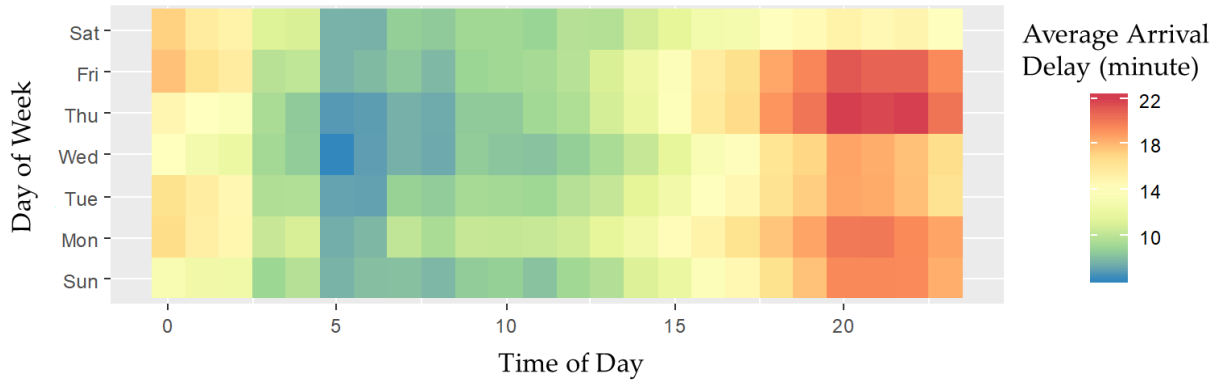


Figure 5. Temporal delay heat map.

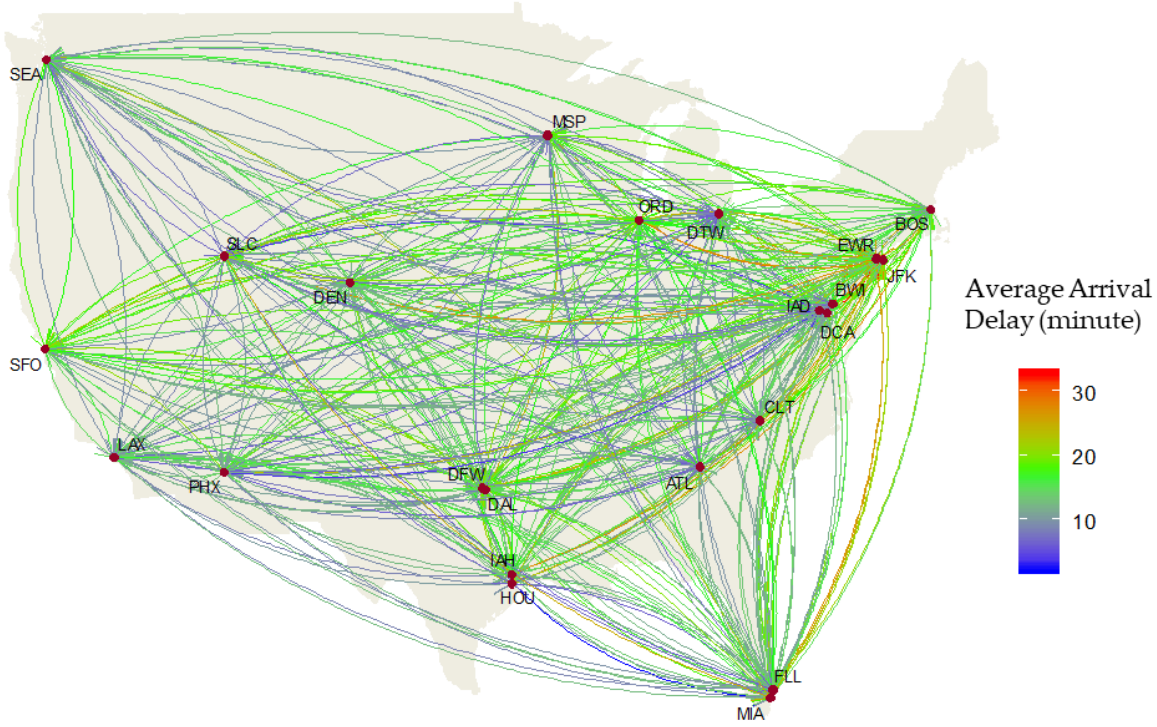


Figure 6. Average arrival delay associated with flights.

Also, all the dataset is imported in Tableau, which enables us to visualize delays of any combination of day, hour, airline, origin, and destination for further analysis. This is very similar to the famous "misery map" in terms of the performance. Figure 7 is a screenshot of this visualization tool which shows delays for top 10 US airports at 2:00 a.m. on January 4, 2018.

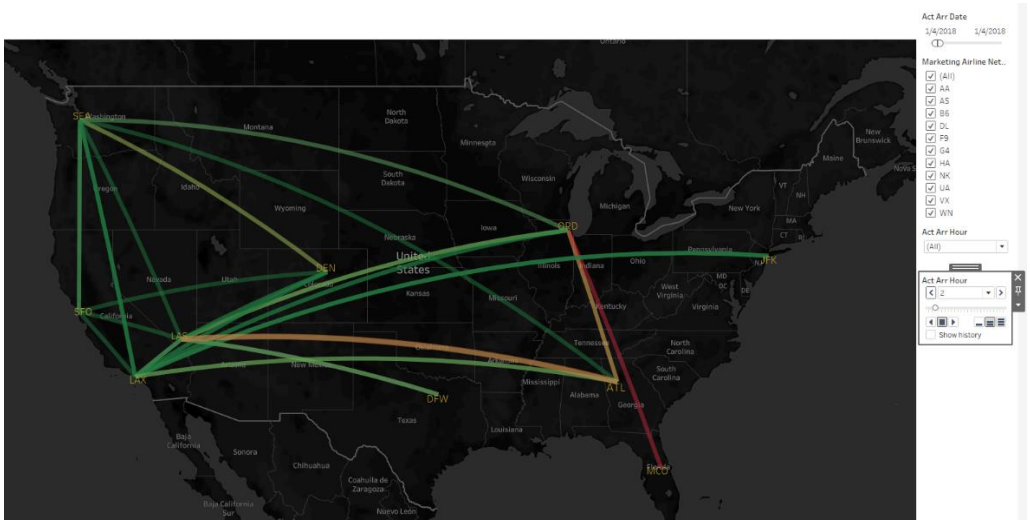


Figure 7. Delays for top 10 US airports at 2:00 am on January 4, 2018.

2.2.2. Market shares of airlines and delay analysis

The market share of airlines in terms of the number of flights is shown in Figure 8. As seen, American and Virgin America are having the highest and the lowest market share, respectively.

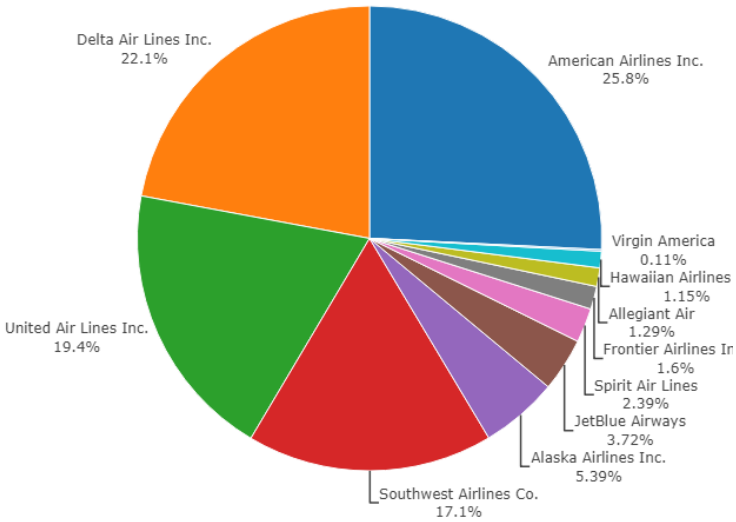
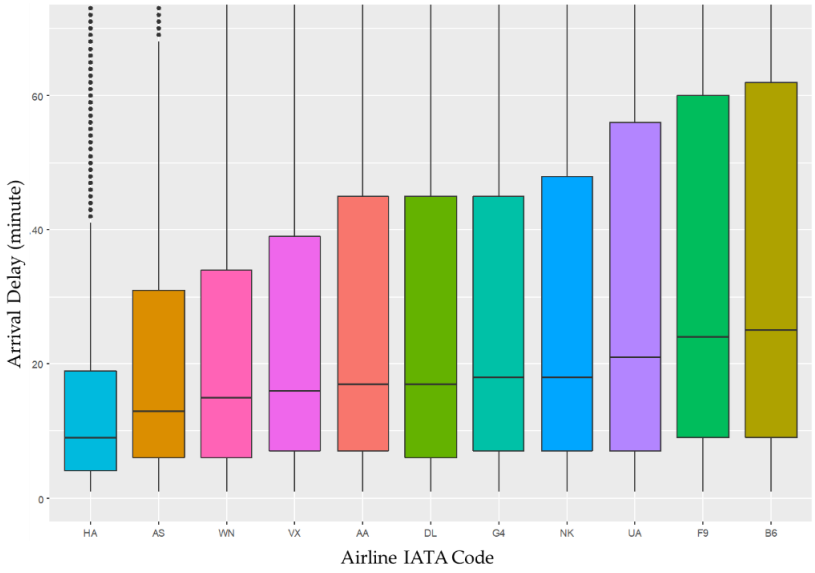


Figure 8. Market share of airlines in terms of the number of flights.

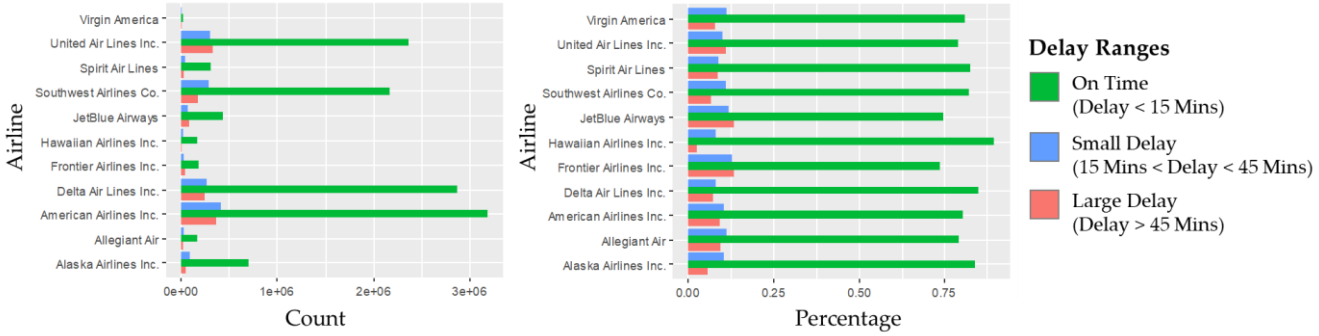
In Figure 9, this chart shows how each airline is contributing to the arrival delay. It should be noted that these boxplots are developed based on only flights that had delay (i.e. flights with zero minute of delay are not considered). It is seen that Hawaiian Airlines has the lowest arrival delay, while JetBlue and Frontier have the most contributions to the arrival delay.



IATA Code	Airline
UA	United Air Lines Inc.
AA	American Airlines Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
HA	Hawaiian Airlines Inc.
VX	Virgin America
G4	Allegiant Air

Figure 9. Boxplots of arrival delay of airlines.

To further investigate how severe the delay of each airline is, Figure 10 is created to show the number and percentage of ranges of flight delays. It is observed that approximately 75 percent of all flights, regardless of their career airlines, arrived on time. While Hawaiian has the lowest rate of large delay-flights, it can be found that JetBlue and Frontier have the highest rate of flights with more than 45 minutes of arrival delay. These results corroborate the earlier conclusion made from Figure 9.



(a) (b)

Figure 10. Delay ranges of airlines: (a) count, and (b) percentage.

2.2.3. Equity index

One of the oldest and most widely used principles of distribution is ranking applicants' rights, which is the "Principle of Proportionality." This principle allocates the resources in proportion to the demand for the resource. In Air Traffic Control (ATC), the mechanism of First-Come/First-Serve is used, which imply the proportionality (Sherry 2010). In Traffic Flow Management (ATFM), the mechanism of First-Scheduled/First-Served is used, whose proportionality is explicit (Sherry 2010). According to the principle of proportionality, airlines should receive delays in proportion to their number of flights scheduled; however, airlines will receive delays consistent with their actual sequences of flights from ATC. The Equity Index is created to describe the difference between scheduled and actual delays (Sherry 2010).

$$\text{Equity Index for Airline (i)} = \frac{\text{The EoDTD for Airline (i)}}{\text{The EoDTD for all Airlines}} \Big/ \frac{\text{Number of Flights for Airline (i)}}{\text{Total Flights for all Airlines}}$$

where EoDTD is end-of-the-day total delay, and i is an airline, 1 through n.

When the proportion of delays is equivalent to the proportion of flights, the equity for the airline is equal to 1. Equity index for an airline less than 1, implies that the airline was allocated delays proportionately less than the number of flights scheduled. This airline benefited from the allocation process. Equity index for an airline greater than 1, implies that the airline was allocated delays proportionately more than the number of flights scheduled. This airline was penalized by the allocation process. In this study, equity indices at noon and at midnight were calculated for each airline. These indices are plotted for each airline, which is shown in Figure 11.

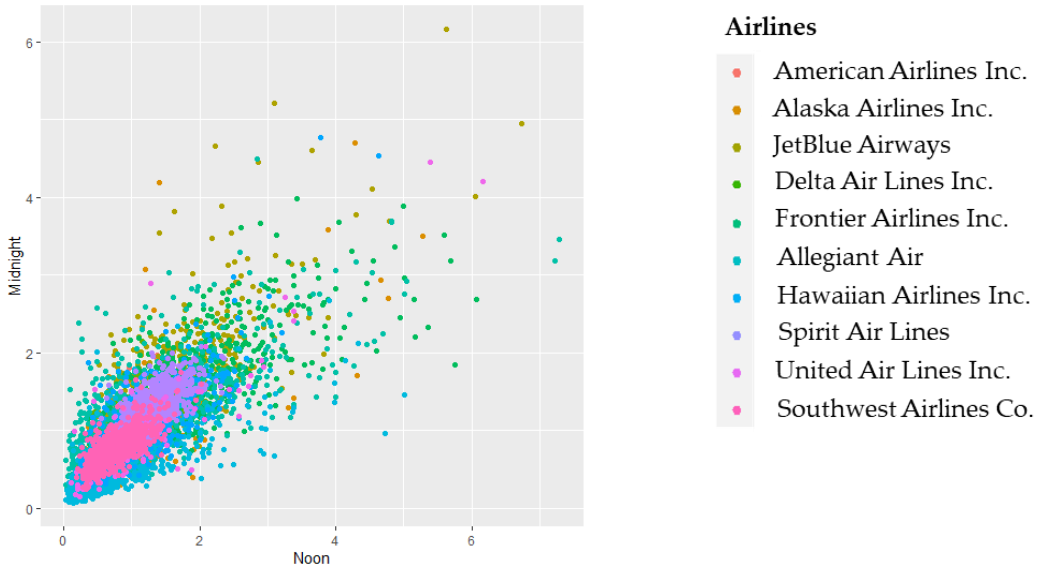


Figure 11. Noon vs midnight equity indices of airlines.

3. Methodology

Methods applied in this study are based upon the objectives, which are twofold. First, it is desired to predict future series of hourly cumulative delays and obtain the value of end-of-the-day delay. To accomplish this objective, time series regression models will be applied. Second, classification models are used to predict the severity class of days based on their cumulative delay at the end of the day. In this study, a novel approach for times series regression analysis is proposed and will be discussed. The following of this section discusses regression and classification approaches applied in this study. For the sake of summarization, cumulative delays are simply referred to as delays, hereafter.

3.1. Time series regression modeling

In this topic, it is attempted to apply time series regression for prediction future hourly delay of an entire air transportation system based on the response history of 12 hours of reported delay and temporal characteristics of the day (e.g. day of the month, whether that day is a holiday, weekend, and etc.). In this study, it is considered that the response history and predicted future are both associated with the same day. That is, the response history is representing the delay of each one-hour time increment from 01:00 a.m. to 12:00 p.m. of a day, which makes the size of the response history being 12. In the following, future responses represent a delay of one-hour time increments for the rest of that day (from 01:00 p.m. to midnight). Besides the 12-hour interval of response history, other different sizes such as 36-hour and 60-hour intervals were tested as well. However, the results were not satisfactory. The size of response history being greater than 12-hour interval means that delays and characteristics of the previous day(s) are taken into account as well. Table 1 shows the data format of our response history (potential explanatory variables) and future response for the purpose of modeling. To simplify, the aim of this section is to predict elements highlighted in red (in Table 1) by utilizing elements highlighted in blue.

Table 1. Data format for time series regression analysis.

Date	Response history[a]				Temporal characteristics				Future response[a]			
	00-01	01-02	…	11-12	DoW[b]	Wk[c]	Season	Holiday	12-13	13-15	…	23-24
2018-01-01	0	0	…	25.88	Mon	No	Winter	Yes	39.63	56.47	…	314.01
2018-01-02	21.13	41.93	…	107.87	Tue	No	Winter	No	125.78	143.09	…	393.88
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019-08-06	50.52	94.01	…	254.96	Sun	Yes	Summer	No	276.80	296.22	…	617.85
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019-12-31	31.89	58.79	…	138.70	Tue	No	Winter	Yes	152.93	169.64	…	306.19

[a] Reported as hourly cumulative delays in ×1000 minute
[b] DoW = day of week
[c] Wk = weekend

3.1.1. Proposed Approach (ATASH):

A novel approach is proposed as a sequential method to predict series of hourly delays within a day. This method is named after the authors, ATASH. Considering the response of a 12-hour time interval (from 01:00 a.m. to 12:00 p.m.) and temporal characteristics as explanatory variables, this method predicts the hourly delay of only the next time increment (hourly delay at 01:00 p.m.) at the first sequence of the method. Prediction results of a regression-type model, which is trained with information of a 12-hour time interval. At the second sequence, the predicted value joins the former set of explanatory variables and the set updates to a 13-hour time interval. Given the new explanatory variables, another regression model, which was trained with information of a 13-hour time interval, applies to predict the value of hourly delay at the next time increment (hourly delay at 02:00 p.m.). This procedure continues until we predict end-of-the-day delay. Logically, the final sequence of the method uses a model trained with a 23-hour time interval. A schematic of method ATASH is shown in Figure 12. In this study, Support Vector Regression (SVR), Artificial Neural Network (ANN), and Random Forest (RF) are utilized separately as the models in the prediction part of the method.

3.1.2. Time series analysis with long short-term memory (LSTM)

In order to take account of the fact that hourly delay is a time series data, a special type of Recurrent Neural Networks (RNNs) called Long Short-Term Memory (LSTM) network is also developed. The LSTM network is capable of capturing the long-term dependencies in the data. To train the LSTM, each day is treated as a data point in the time series dataset. As discussed previously, the input of the model is a sequence of delays up to 12:00 p.m. and the target sequence to be estimated is a sequence of the delay for the rest of the day. Note that the LSTM model is capable of predicting a sequence of hourly delay for the rest of the day until midnight given a sequence of hourly delay up to noon. Thus, opposite to the sequential modeling approach, as described in the previous section, it is not necessary to create recursive sequences. This recursive process is taken care of by LSTM automatically.

3.1.3. Measures of fit

To examine the performance of time series regression methods, Root Mean Square Error (RMSE) formulated as Eq. (1) is selected as the primary criterion for each method. The RMSE indicates the variability between predicted and actual hourly delays.

$$RMSE = \sqrt{\frac{1}{mn}\sum_{\forall i,j}(\hat{y}_{ij} - y_{ij})^2} \quad (1)$$

where

$\hat{y}_{ij}$ = predicated hourly delay of time interval i at day j (×1000 minute),

y_{ij} = actual hourly delay of time interval i at day j (×1000 minute),

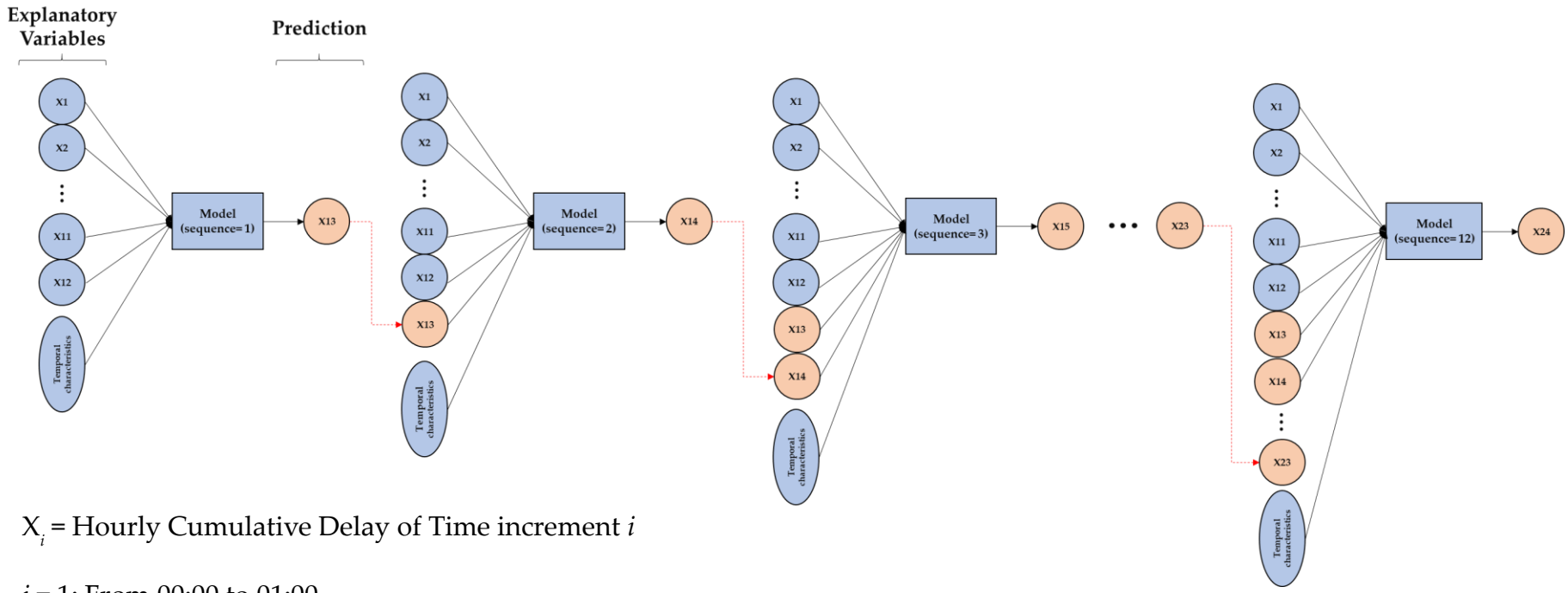
i = ith one-hour interval of future response ($1 \leq i \leq m$),

j = jth day of case study period ($1 \leq j \leq n$),

m = number of one-hour intervals of future response, also referred to as future response size (decided to be equal to 12 in this study),

n = number of days within the validation dataset (=365 days in this study).

Obviously, a smaller value of RMSE denotes a better fit of the model to the data. That being said, parameters of the trained regression models (e.g. cost parameter in SVR, number of hidden layers in ANN, number of trees in RF, number of units in LSTM) were tuned and selected to obtain the lowest RMSE.



X_i = Hourly Cumulative Delay of Time increment i

i = 1; From 00:00 to 01:00

i = 2; From 01:00 to 02:00

⋮

i = 24; From 23:00 to 24:00

Temporal Characteristics include DoW, Wk, Season, and Holiday

Figure 12. Schematic of the proposed sequential method for time series regression modeling.

3.2. Classification modeling

In the following of time series regression, this study attempts to apply classification methods for predicting the class of end-of-the-day delay severity. This classification is based on information extracted from a dataset consisting of records of delay of each one-hour interval with known classes of end-of-the-day delay severity. In the beginning, classification methods are first applied to predict delay class (at the end of the day) of each day based on hourly delay observations up to 12:00 p.m. of that day. That is, knowing the hourly delay from 01:00 a.m. to 12:00 p.m. of a day, we predict the delay class of that day. Obviously, if that interval of delay information extends to a further time, such as 01:00 p.m. or 02:00 p.m., a more accurate prediction would be the result. Table 2 shows the required data format for classification modeling. Mathematically speaking, the potential exoplanetary variables of classification modeling can be a combination of any hourly delay and/ or temporal characteristics. Potential explanatory variables and class prediction are highlighted in blue and red, respectively. The value of a model lies on utilizing less information to obtain the highest accuracy. That being said, using delay information of up to any time after 12:00 p.m. will give a more accurate model, while it degrades the value of the model. In this regard, the hourly delay associated with time increments after noon are illustrated with lighter shades. The latter, the lighter. Additionally, classes reported in this table were determined based on the value of end-of-the-day delay. The details of classification methods are discussed later.

Table 2. Data format for classification modeling.

Date	Hourly cumulative delay[a]								Temporal characteristics				Class[d]
	00-01	01-02	…	11-12	12-13	13-15	…	23-24	DoW[b]	Wk[c]	Season	Holiday	
2018-01-01	0	0	…	25.88	39.63	56.47	…	314.01	Mon	No	Winter	Yes	Normal
2018-01-02	21.13	41.93	…	107.87	125.78	143.09	…	393.88	Tue	No	Winter	No	Semi-Normal
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019-08-06	50.52	94.01	…	254.96	276.80	296.22	…	617.85	Thu	No	Summer	No	Abnormal
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019-12-31	31.89	58.79	…	138.70	152.93	169.64	…	306.19	Tue	No	Winter	Yes	Normal

[a] Reported as hourly cumulative delays in ×1000 minute
[b] DoW = day of week
[c] Wk = weekend
[d] Classified based on the mean of cumulative end-of-the-day delay (293.81 e03 minutes):
Normal: Delay ≤ 1.15 × Mean of Delay at End of the Day
Semi-Normal: 1.15 × Mean of Delay at End of the Day < Delay ≤ 1.85 × Mean of Delay at End of the Day
Abnormal: Delay > 1.85 × Mean of Delay at End of the Day

3.2.1. Prediction Accuracy

To examine the performance of the classification prediction model with S levels of delay severity class, prediction results of each model can be summarized in a confusion matrix (Stehman 1991) as shown in Table 3. In this confusion matrix, p_{ij} shows the number of days associated with delay severity level i, predicted as severity level j. Also, r_{ij} is the ratio of days associated with delay severity level i, predicted as severity level j. In other words, r_{ij}=p_{ij}/N_i, where N_i is the actual number of days with the delay of severity level i in the validation dataset. In this regard, the value of r_{ii} shows the correct prediction ratio of days with the delay of severity level i, it is also known as the detection rate of class i. The metric to measure the overall performance of a model is the overall correct prediction rate, also known as the accuracy of the model and $Acc_{total} = \sum_{i=1}^{S} p_{ii} / \sum_{i=1}^{S} N_i$.

Table 3. Confusion Matrix of a delay severity classification method.

	Delay Severity Level	Predicted				Actual Number of Days of Delay Level
		1	2	…	S	
Actual	1	p_{11} r_{11}	p_{12} r_{12}	…	p_{1S} r_{1S}	N_1
	2	p_{21} r_{21}	p_{22} r_{22}	…	p_{2S} r_{2S}	N_2
	⋮	⋮	⋮	⋮	⋮	⋮
	S	p_{S1} r_{S1}	p_{S2} r_{S2}	…	p_{SS} r_{SS}	N_S

3.3. Prediction Models

In the fulfillment of the objectives of both regression and classification approaches, popular machine learning models including Support Vector Regression/ Machine (SVR/ SVM), Artificial Neural Network (ANN), and Random Forests were applied. Added to machine learning models, a well-known deep learning model, Long Short-Term Memory (LSTM), was used as well. The performance of these models was tested and compared to each other. The following of this section presents these models and discusses a brief introduction to each of them.

3.3.1. Artificial Neural Network

Artificial Neural Network (ANN) is a non-linear data modeling tool. ANN is used to simulate a complicated relationship between a set of inputs and outputs without any prior assumptions or any available mathematical relationships between the inputs and outputs. ANN includes a group of interconnected artificial neurons to express the variability and fluctuations related to the datasets. To be specific, ANN consists of a

multi-layered structure of neurons, each with an input (its input is the output of the previous layer of neurons) and an output. The strength of each connection between the neurons is represented by a weight factor.

3.3.2. Support Vector Machine/ Regression

Support Vector Machine/ Regression (SVM/ SVR) is a classification/ regression tool, which is a supervised machine learning algorithm. The main idea of the SVM/SVR method is to generate the optimal separating hyperplane, which can divide the input variable space into two subspaces and. All data points located in one of the subspaces are regarded as with the same class. The distance between the hyperplane and the closest data points of each class is called the margin. The SVM method is to figure out the optimal separating hyperplane via maximizing the margin. Readers can find more details on this method in (Han, Pei and Kamber 2011).

3.3.3. Random Forest

Random Forest (RF) is a classification/ regression model that uses the same if-then-else rule as with the Decision Trees method to generate a class for an input. RF is an ensemble learning method, which is composed of many decision trees. During the model training period, every decision tree in RF is trained by the independent identically distributed random vectors. As a classification, all the decision trees in this RF generate a class for the new input and then the most popular class of these generated classes is the final class result of the RF. More details on RF can be found in (Breiman 2001).

3.3.4. Long Short-Term Memory

Long Short-Term Memory (LSTM) is a specific type of Recurrent Neural Network (RNN) that can handle time-series input. In RNNs, each time the previous output is taken to the next hidden layer and trained together, but the short-term memory has a larger impact on the next hidden layer than long-term memory. LSTM overcome this defect of RNN, which can select and reserve the important information of the previous layers and then take it to the next hidden layer.

3.4. Converting numeric data to categories

For the purpose of classification modeling, the response variable should be in the form of categorical variable. However, in this study, the response variable is the end-of-the-day delay, which is a numeric variable. This study attempts to convert continuous values into categories. Generally, there are no certain rules for converting numeric data into categories. This section presents two methods of categorizing data.

3.4.1. Categorizing data with clustering

This approach uses a clustering algorithm to divide data into categories with similar characteristics. This method is more efficient in problems with multiple distinct variables. There are several methods of clustering, of which K-Mean is selected as a method to categorize our dataset. The first step is to create the histogram of end-of-the-day delay, as shown in Figure 13.

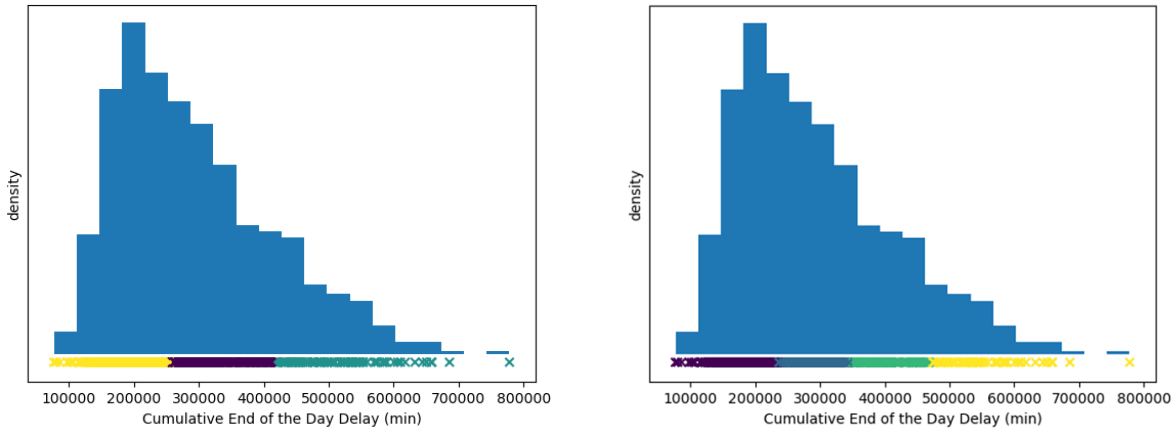


Figure 13. Density histogram of end-of-the-day delay with rugs showing: (a) 3 classes, and (b) 4 classes.

Since the data does not contain any labels, i.e. there is no information whether a specific value of delay is considered a normal day or not, an unsupervised K-Mean clustering algorithm is used to determine the number of clusters to categorize the end-of-the-day delay. Figure 14 shows how the number of clusters changes the value of within-cluster sum-of-squares criterion. By performing the elbow analysis which is a typical procedure to find the optimal number of clusters, 3 and 4 seem to be two good choices.

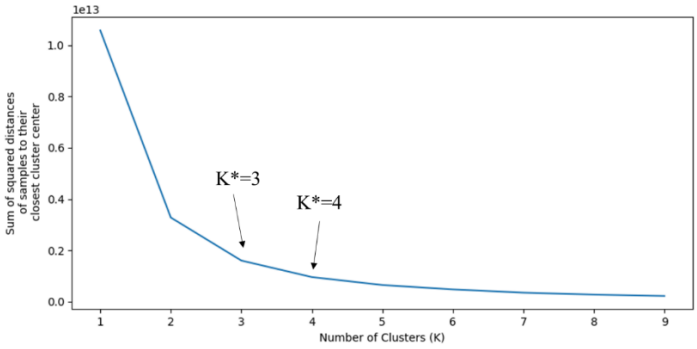
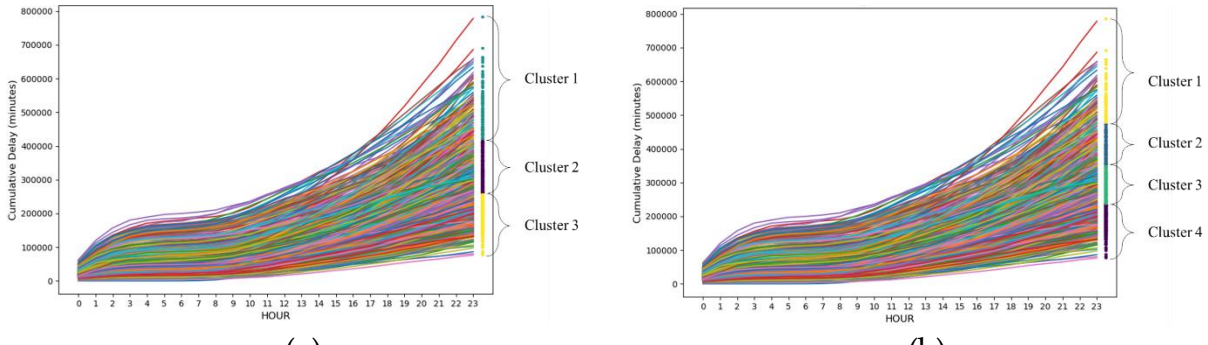


Figure 14. Sum-of-squares vs number of clusters.

To give a visual example of how we converted end-of-the-day delay to categories, Figure 15 is shown for 3 and 4 clusters. This approach has a problem of not representing for real-world severity of classes. In other words, in a case of 3 clusters, most records of clusters number 2 and 3 which are referred to as normal and semi-normal conditions are below the mean of end-of-the-day delay. That being said, it was concluded to apply another approach for categorizing the data.



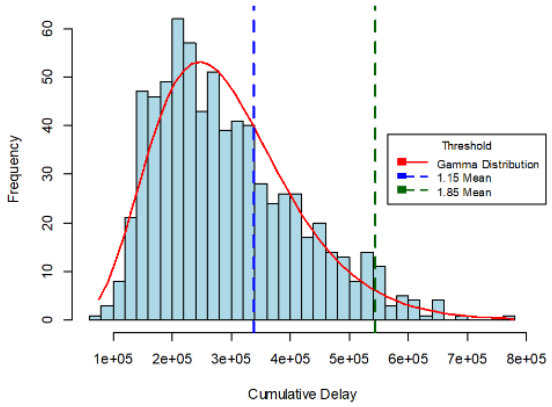
(a) (b)

Figure 15. Conversion of end-of-the-day delay to: (a) 3 categories, and (b) 4 categories.

3.4.2. Categorizing data by a range of values

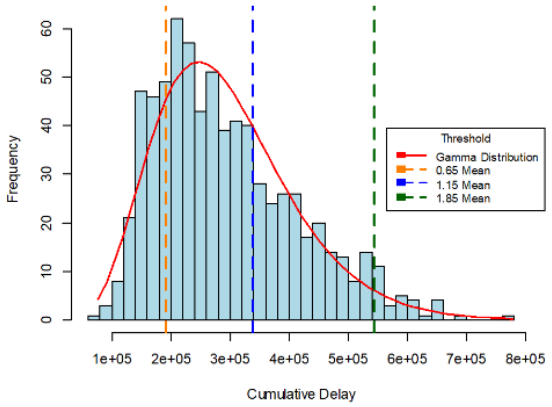
Another approach is to create classes based on logical cut-off values. In this study, cut-off values are determined according to the mean of end-of-the-day delay. To this aim, a histogram of end-of-the-day delay was created. It is worth mentioning that a Gamma distribution (shape parameter = 6.255, rate parameter = 0.0213) was fitted to end-of-the-day delay histogram. Then, sets of cut-off values were applied to histograms as shown in Figure 16. Candidate scenarios of categorization were determined as categorizing data with 3, 4, and 5 classes. Thresholds associated with each criteria of categorization are listed in Figure 16 as well. Figure 16 also shows histograms of end-of-the-day delay for each candidate number of classes.

Performance of different classification models (in terms of overall prediction accuracy) was tested for each candidate scenario of classification. Results, reported in Table 4, indicate that categorizing the end-of-the-day delay into 3 classes: 1. Normal, 2. Semi-Normal, and 3. Abnormal gives the highest accuracy. Therefore, it is decided to have 3 classes of delay for the purpose of analytics and modeling.



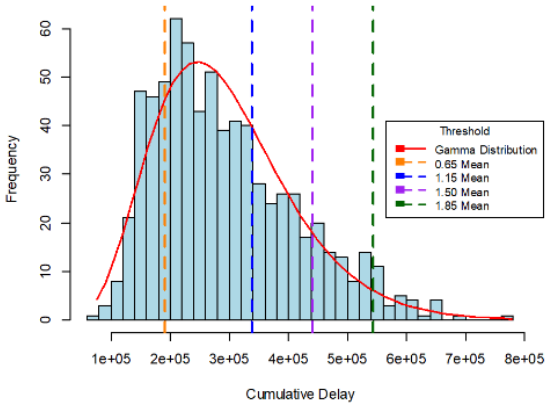
Normal: Delay≤1.15×Mean
Semi-Normal: 1.15×Mean<Delay≤1.85×Mean
Abnormal: Delay>1.85×Mean

(a)



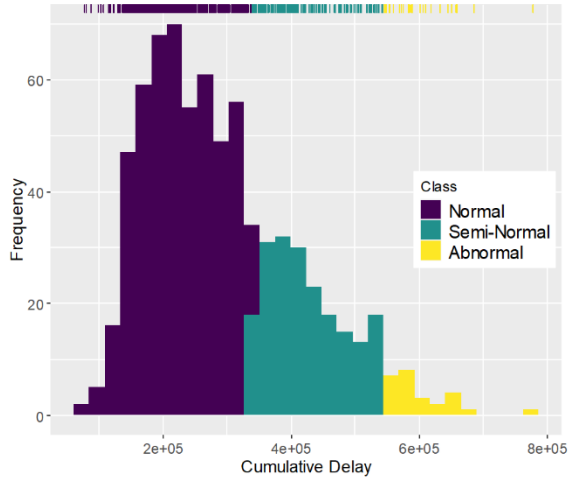
Good: Delay≤0.65×Mean
Normal: 0.65×Mean<Delay≤1.15×Mean
Semi-Normal: 1.15×Mean<Delay≤1.85×Mean
Abnormal: Delay>1.85×Mean

(b)

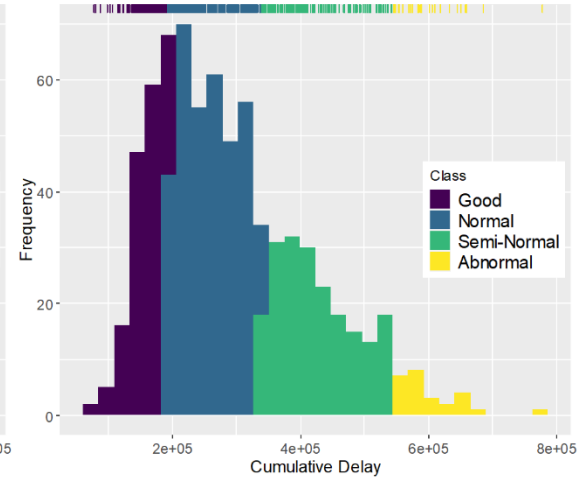


Good: Delay≤0.65×Mean
Normal: 0.65×Mean<Delay≤1.15×Mean
Semi-Normal: 1.15×Mean<Delay≤1.50×Mean
Semi-Abnormal: 1.50×Mean<Delay≤1.85×Mean
Abnormal: Delay>1.85×Mean

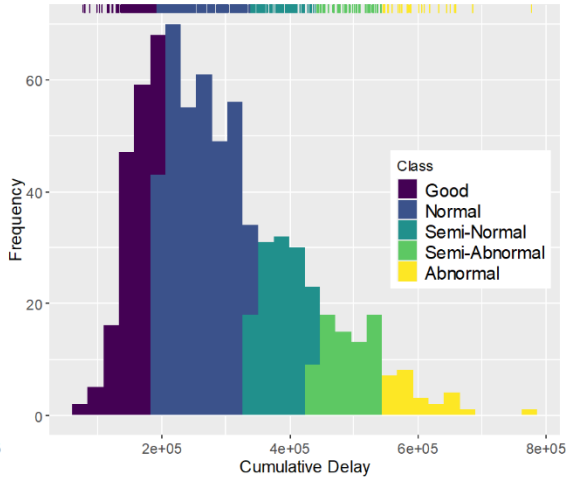
(c)



(1)



(2)



(3)

Figure 16. Candidate scenarios of categorization: (a) 3 classes, (b) 4 classes, and (c) 5 classes. Histograms of (1) 3 classes, (2) 4 classes, and (3) 5 classes

Table 4. Overall accuracy comparison measures.

Number of Classes	Size of History Response	Overall Accuracy		
		Prediction Method		
		SVM	RF	LSTM
3 Classes	12	77.20%	73.35%	78.02%
	13	79.67%	75.82%	79.12%
	14	80.77%	78.85%	79.67%
	15	83.52%	80.77%	80.49%
4 Classes	12	67.03%	65.11%	61.26%
	13	69.51%	67.31%	62.91%
	14	71.98%	68.68%	68.41%
	15	73.63%	71.15%	70.60%
5 Classes	12	56.04%	51.92%	54.40%
	13	60.71%	55.49%	57.42%
	14	61.81%	60.16%	60.71%
	15	66.76%	60.99%	61.81%

3.5. Data Balancing

In this study, it is aimed to identify the condition of a day based on its end-of-the-day delay class. Having 2 years of data of flights, the total number of extracted data would be 728 days (March 11th, 2018 and March 10th, 2019 were removed from data records). Given categorizing data into 3 classes, the total number of Abnormal days observed is 26 days. This number is about 3.6 percent of the entire dataset, making this dataset highly imbalanced. Table 5 reports the total number of each class records. If the structure of the dataset retains imbalanced, it may cause issues of misclassification and low accuracy. To avoid such problems, the dataset got balanced.

There are different techniques of data balancing, among which the Synthetic Minority Over-sampling Technique (SMOTE) was selected as the data balancing technique of this study. SMOTE, introduced by (Chawla, et al. 2002), creates synthetic data points of minority classes (Abnormal and Semi-Normal in our case) with respect to their *k* (which assigned to 5 in our study) nearest neighbors. Figure 17 shows the schematic of SMOTE. For more complete information, readers are suggested to read (Chawla, et al. 2002).

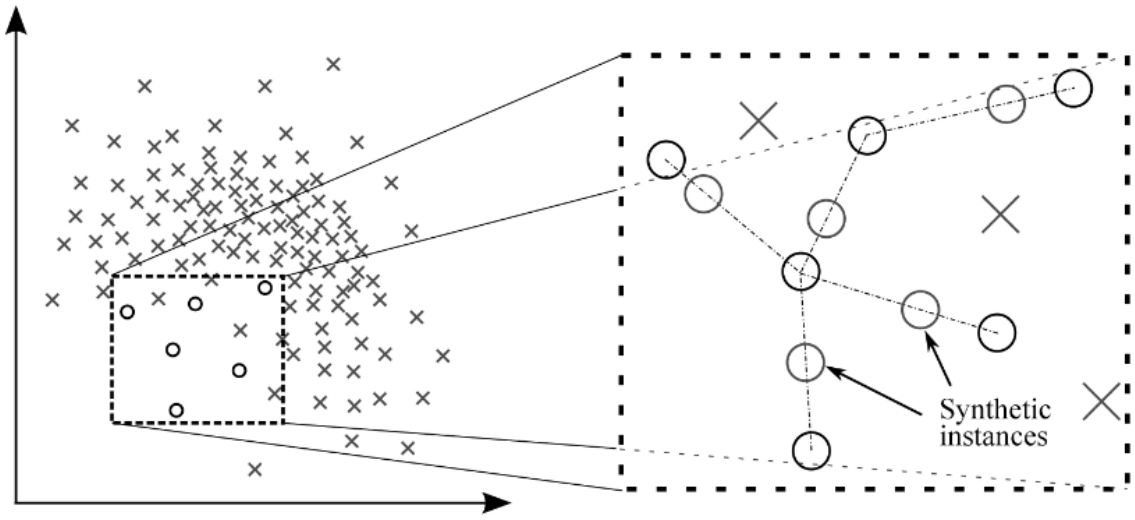


Figure 17. Schematic of SMOTE (source: (Walimbe 2017)).

The imbalanced dataset split into train (including 364 days for year 2018) and test (including 364 days for the year 2019) datasets. Applying SMOTE, a balanced train dataset based on the original imbalanced train dataset is generated. It was tried to have approximately equal numbers of each class in the trained dataset. Table 5 shows the number of records of each class in train, test, and balanced train datasets.

Table 5. Classes of delays in dataset.

Delay Class	Imbalanced Train	Imbalanced Test	Imbalanced Total	Balanced Train
Normal	270 (74.18%)	234 (64.29%)	504 (69.23%)	270 (34.62%)
Semi-Normal	82 (22.53%)	116 (31.87%)	198 (27.20%)	246 (31.54%)
Abnormal	12 (3.29%)	14 (3.85%)	26 (3.57%)	264 (33.85%)
Overall	364	364	728	780

4. Results

This section discusses the results obtained from both regression and classification modeling.

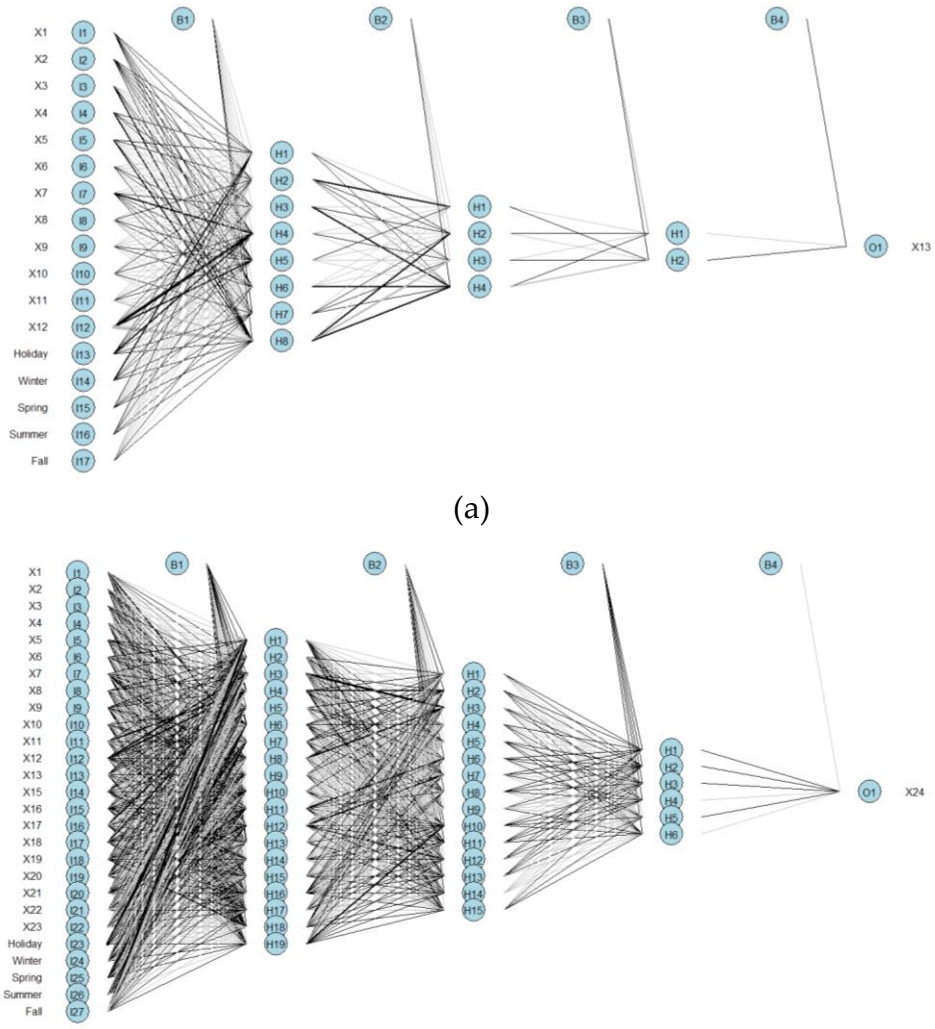
4.1. Time series regression modeling results

Results obtained from approach ATASH are reported in Table 6. Results show the reported RMSE of regression models used in the prediction part of this method.

Table 6. Results of approach ATASH.

Model Used	RMSE (minute)
SVR	29111.54
ANN	28468.69
RFF	31547.25

For training the SVR model, the values of parameters were determined as *Cost* = 2.5 and *Gamma* = 5. Training RF needs tuning of the number of trees to grow (ntree) and the number of variables randomly sampled as candidates at each split (mtry). It found that the best result obtains when *ntree* = 200. The default value of *mtry* is suggested in the literature as the square root of the number of explanatory variables (we have 17 to 28 variables based on the sequence number of the method). Thus, *mtry* = 4 and 5 were chosen. For training the ANN, 3 hidden layers were chosen, and the number of neurons on each hidden layer depends on the sequence number of the model. To bring examples, there are 8, 4, and 2 neurons on hidden layers in the first sequence of the model (predicting X13). While, there are 19, 15, and 6 neurons on hidden layers at the last sequence (predicting X24). This example is shown in Figure 18.



(a)

(b)

Figure 18. Schematic of the used ANN at: (a) the first and (b) the last sequences of approach ATASH.

To visualize the performance of approach ATASH, delay profiles of four sample days are shown in Figure 19. Note that the predicted values are based on the results obtained from the ANN model.

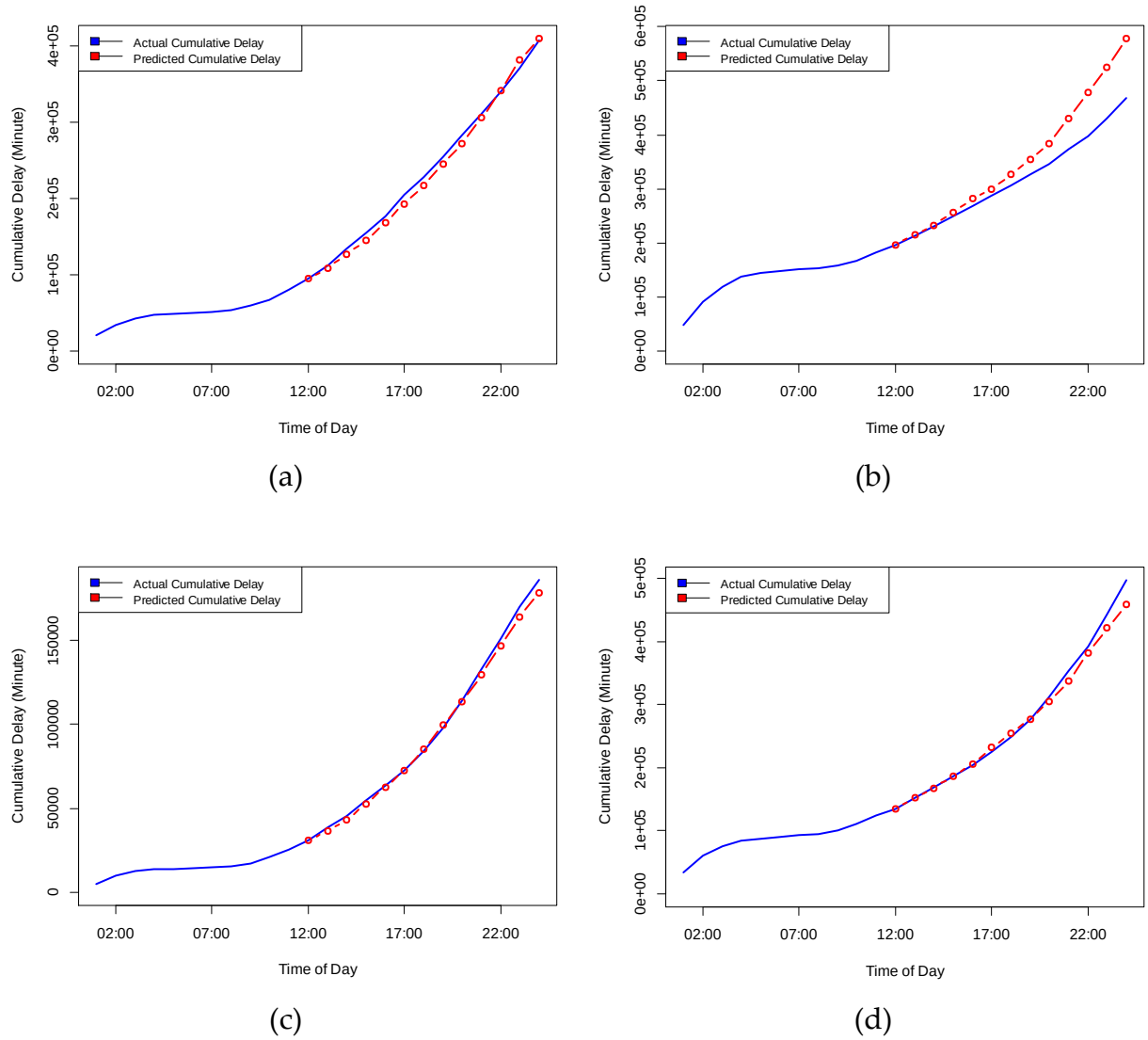


Figure 19. ANN predicted delay profiles of: (a) Mar 1st, (b) Jul 22nd, (c) Sep 26th, and (d) Dec 3rd, 2019.

Additionally, results from LSTM models with different time of prediction from noon to 3:00 p.m. (history size varies from 12 to 15) are reported for NAS and some sample airlines. The results of LSTM models can be found in Table 7. Comparing the results of LSTM (12) and ANN from approach ATASH (the best result of approach ATASH), LSTM performance is slightly better than ATASH. That is, it was decided not to use ATASH for the purpose of time series regression for different airlines and different time of prediction.

Table 7. Results of LSTM time series regression.

Prediction Method*	RMSE (minute)				
	NAS	Airline			
		American	JetBlue	United	Delta
LSTM (12)	27907.97	9439.43	2988.1	10099.01	10908
LSTM (13)	25662.93	8318.17	2916.1	9018.26	9644.91
LSTM (14)	24008.99	8147.22	2796.5	8990.05	8824.65
LSTM (15)	24126.56	8007.88	2657.3	7799.35	9095.34

* Numbers in parentheses account for the size of response history.

To give an illustrative compression of LSTM and ATASH, delay profiles predicted by LSTM (12) were created for the exact same days that were discussed earlier. These profiles are shown in Figure 20.

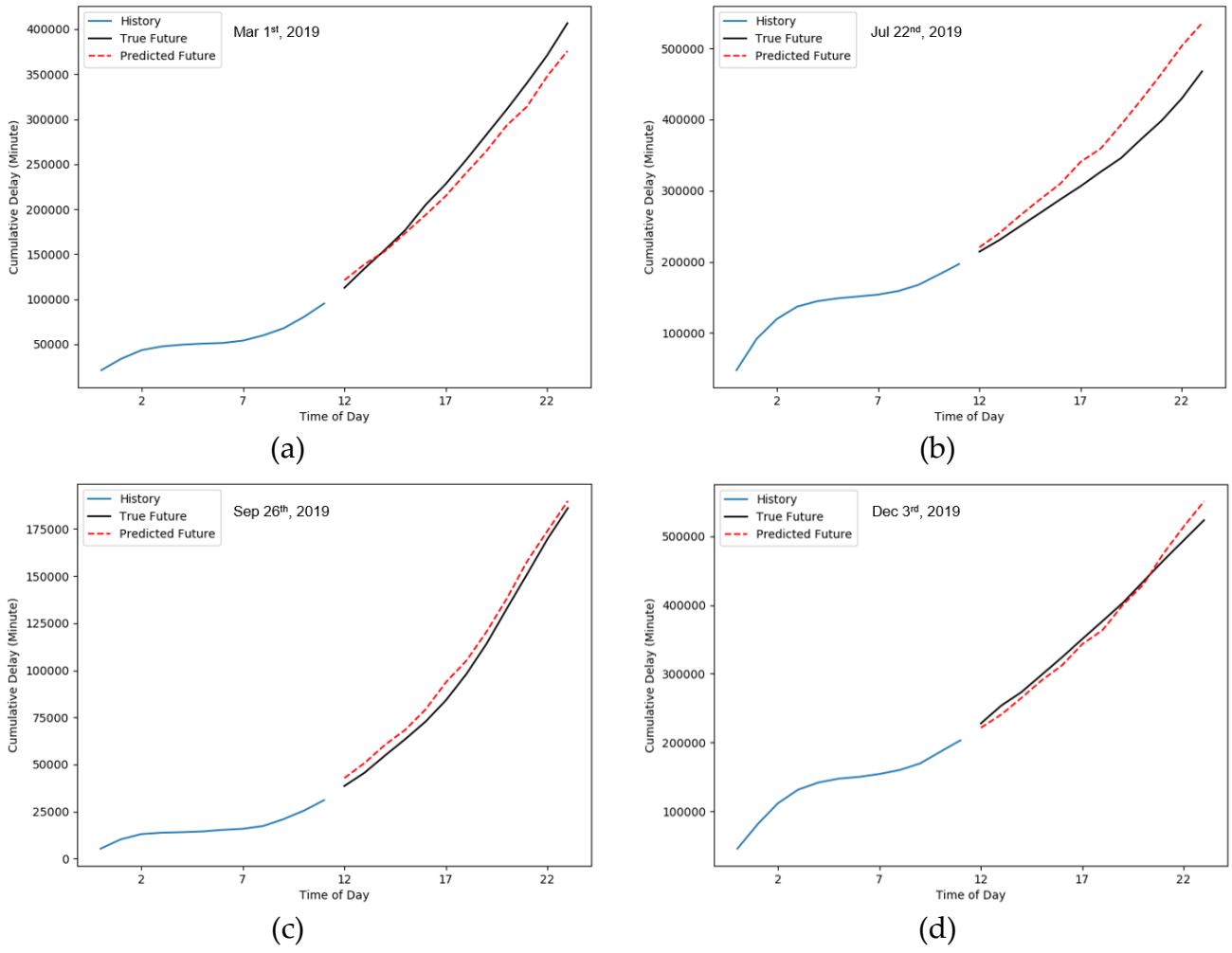


Figure 20. LSTM (12) predicted delay profiles of: (a) Mar 1st, (b) Jul 22nd, (c) Sep 26th, and (d) Dec 3rd, 2019.

4.2. Classification modeling results

Three prediction models were estimated using the training datasets with response history sizes of from 12 to 15 hours. The test dataset was used for assessing the methods' prediction accuracy and detection rates of classes. For training the SVM model, the values of parameters were determined as *Cost* = 2 and *Gamma* = 5. In RF, parameters were determined as *ntree* = 100 and *mtry* = 4.

As it was discussed, each prediction method was executed on a balanced dataset with different sizes of history response from 12 to 15 hours. Given three prediction methods and four sizes of history response, twelve sets of result are expected. These twelve sets are reported in a confusion matrix showing the prediction counts and detection rates for each delay class. This confusion matrix is shown in Table 8. Also, prediction overall accuracy rates are reported in Table 9.

Table 8. Confusion matrix for delay severity prediction models with different size of history response.

Delay Severity Level	Prediction Method*	Classified p_{ij}			r_{ij}		
		Normal	Semi-Normal	Abnormal	Normal	Semi-Normal	Abnormal
Normal N_1= 234	SVM (12)	188	46	0	**80.34%**	19.66%	00.00%
	SVM (13)	195	39	0	**83.33%**	16.67%	00.00%
	SVM (14)	201	33	0	**85.90%**	14.10%	00.00%
	SVM (15)	206	28	0	**88.03%**	11.97%	00.00%
	RF (12)	184	50	0	**78.63%**	21.37%	00.00%
	RF (13)	189	44	1	**80.77%**	18.80%	00.43%
	RF (14)	195	38	1	**83.33%**	16.24%	00.43%
	RF (15)	197	36	1	**84.19%**	15.38%	00.43%
	LSTM (12)	212	22	0	**90.60%**	9.40%	00.00%
	LSTM (13)	212	22	0	**90.60%**	9.40%	00.00%
	LSTM (14)	210	24	0	**89.74%**	10.26%	00.00%
	LSTM (15)	212	22	0	**90.60%**	9.40%	00.00%
Semi-Normal N_2= 116	SVM (12)	15	82	19	12.93%	**70.69 %**	16.38%
	SVM (13)	14	84	18	12.07%	**72.41%**	15.52%
	SVM (14)	15	82	19	12.93%	**70.69%**	16.38%
	SVM (15)	9	87	20	7.76%	**75.00%**	17.24%
	RF (12)	21	72	23	18.10%	**62.07%**	19.83%
	RF (13)	20	76	20	17.24%	**65.52%**	17.24%
	RF (14)	14	81	21	12.07%	**69.83%**	18.10%
	RF (15)	10	86	20	8.62%	**74.14%**	17.24%
	LSTM (12)	47	66	3	40.52%	**56.90%**	2.59%
	LSTM (13)	43	70	3	37.07%	**60.34%**	2.59%
	LSTM (14)	36	74	6	31.03%	**63.79%**	5.17%
	LSTM (15)	35	75	6	30.17%	**64.66%**	5.17%
Abnormal N_3= 14	SVM (12)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (13)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (14)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (15)	0	3	11	00.00%	21.43%	**78.57%**
	RF (12)	1	2	11	7.14%	14.29%	**78.57%**
	RF (13)	0	3	11	00.00%	21.43%	**78.57%**
	RF (14)	0	3	11	00.00%	21.43%	**78.57%**
	RF (15)	0	3	11	00.00%	21.43%	**78.57%**
	LSTM (12)	1	7	6	7.14%	50.00%	**42.86%**
	LSTM (13)	0	8	6	0.00%	57.14%	**42.86%**
	LSTM (14)	0	8	6	0.00%	57.14%	**42.86%**
	LSTM (15)	0	8	6	0.00%	57.14%	**42.86%**

* Numbers in parentheses account for the size of response history.

Table 9. Prediction accuracy rates.

Size of History Response	Overall Accuracy		
	Prediction Method		
	SVM	RF	LSTM
12	77.20%	73.35%	78.02%
13	79.67%	75.82%	79.12%
14	80.77%	78.85%	79.67%
15	83.52%	80.77%	80.49%

To give an illustration of how these models work in general, delay profiles of each classified level are shown in Figure 21. In this figure, classes were determined based on the results obtained from SVM (12) in Table 8.

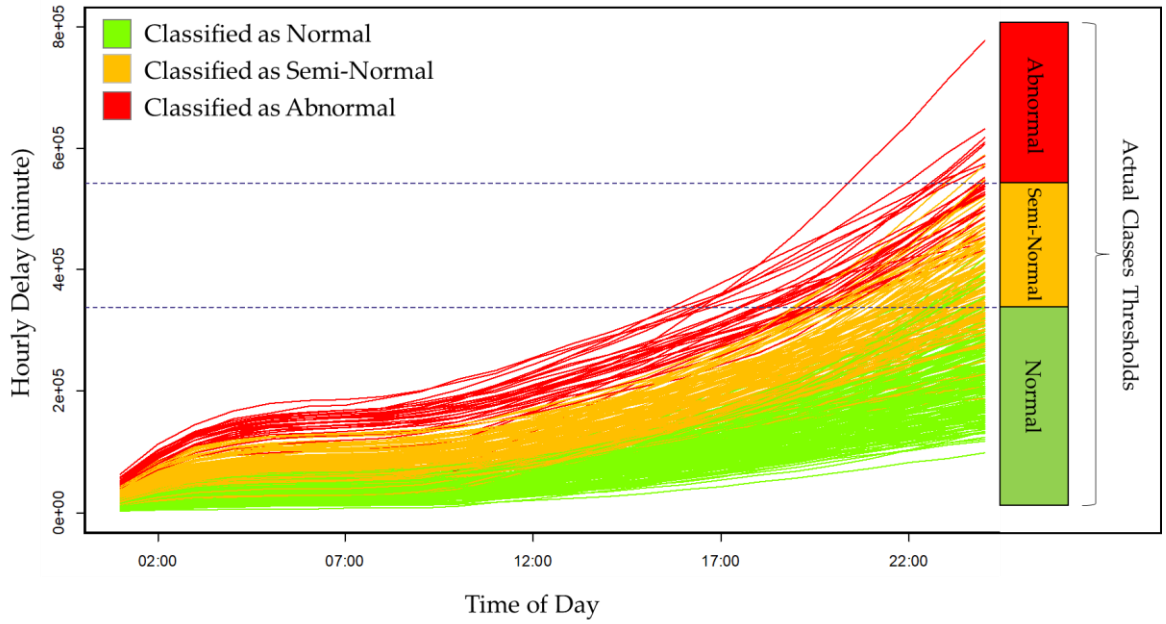


Figure 21. Classified vs actual delay profiles.

4.2.1. Equity index analytics

This section investigates the potential for predicting the class of delay through the analytics of equity indices. For this purpose, equity indices associated with noon and midnights are plotted based on airline and class of delay. Results, which are shown in Figure 22, do not indicate any specific pattern of indices based on the class of delays. Note that all delay classes used in this section are determined based on results of SVM (12).

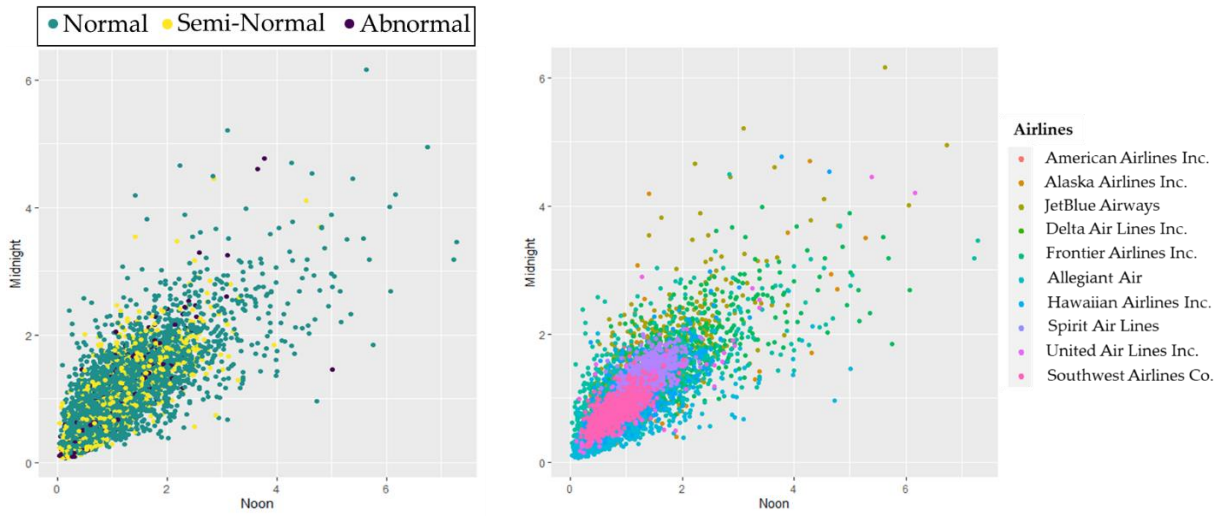


Figure 22. Equity index plot based on: (a) classes of delay, and (b) airlines.

For further analysis, equity index plots were developed for major airlines: American, United, Delta, and Southwest. They have the highest market shares with about 85% in total. Plots can be found in Figure 23. The diagonal line means that equity indices of Noon and Midnight are equal. Points below the diagonal line (blue) mean that the equity index at noon of a given day is greater than that value at midnight of that day (i.e. equity index ratio of noon to midnight is greater than 1). Similarly, points above the diagonal line (red) are showing that the equity index at midnight was higher than noon (equity index ratio of noon to midnight is less than 1).

In most cases, the equity index gets worse from noon to midnight for American. It means that the performance of American is getting worse as time goes on. But it has nothing with the class of delay and no pattern observed. The same thing applies to United, Delta, and Southwest. Only it can be seen the number of points above and below the diagonal line are relatively equal for these airlines.

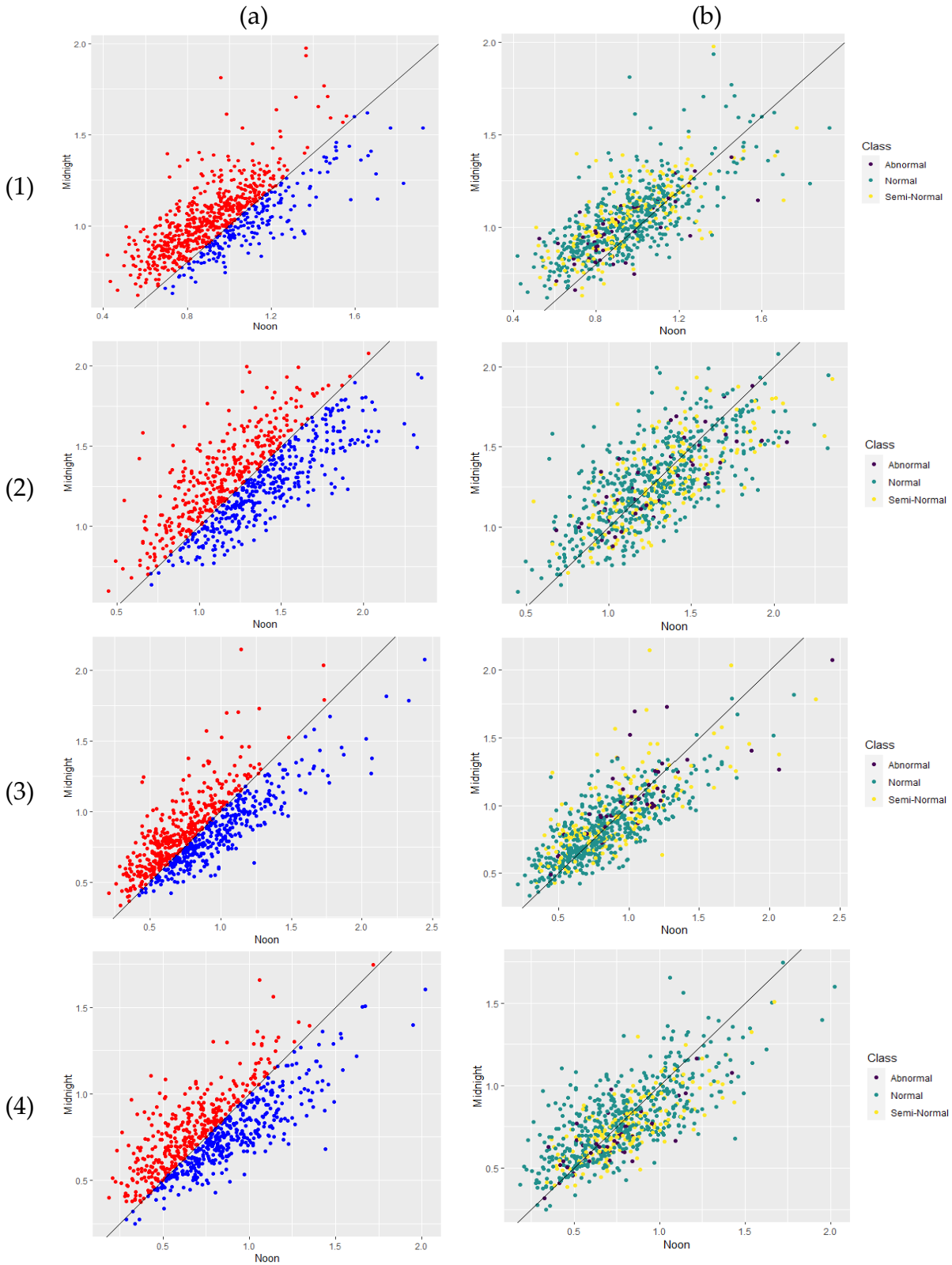


Figure 23. Equity index plots based on: (a) equity index ratio, and (b) airlines: (1) American, (2) United, (3) Delta, and (4) Southwest.

The potential for predicting delay classes based on the equity index at noon was also investigated. To do so, the parallel coordinate plot of equity indices at noon of discussed airlines were developed based on the classes of delay. This plot is shown in Figure 24. Visually speaking, there is no correlation between equity indices at noon and classes of delay.

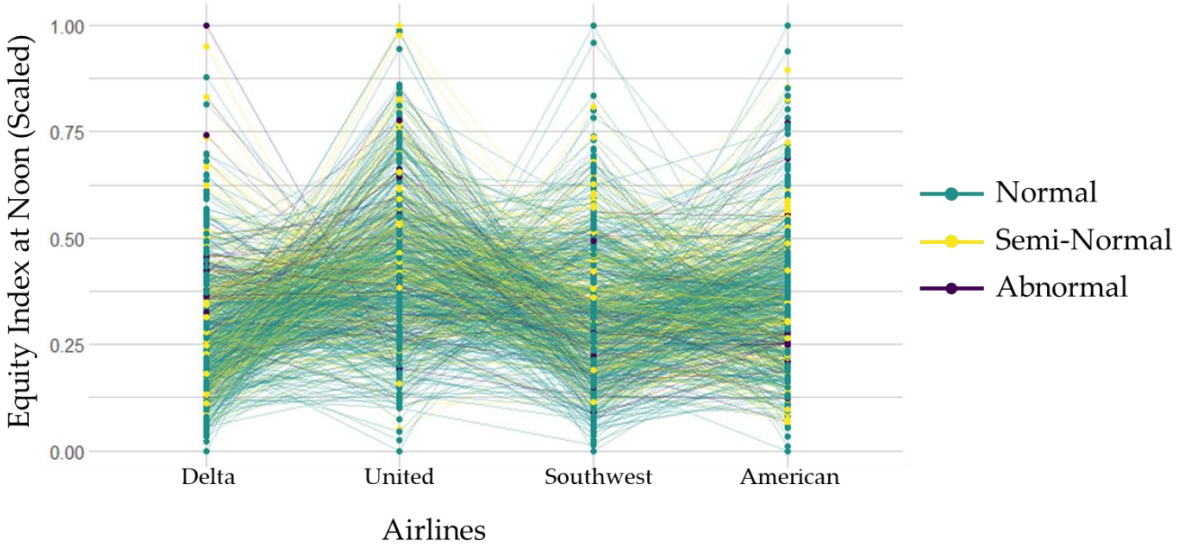


Figure 24. Parallel coordinate plot of equity index.

To statistically investigate the presence of any correlation, a correlation plot was created and is shown in Figure 25. The highest value of correlation is related to the correlation of Abnormal class and Delta, which is 0.17. It can be concluded that there is almost no correlation between equity indices of airlines and classes of delay.

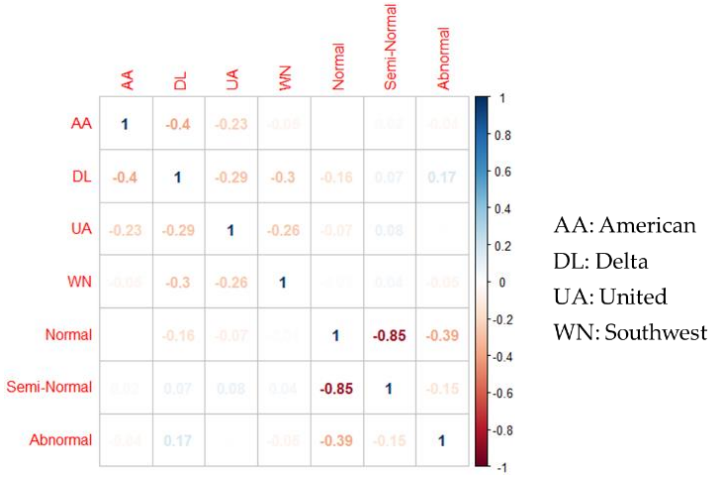


Figure 25. Correlation plot of airlines equity index and classes of delay.

4.2.2. Misclassified days

The purpose of this study mostly focuses on abnormal days. As of that, days misclassified as abnormal or abnormal days being misclassified are subject to analyze more deeply. As a result of SVM (12) is seen in the confusion matrix, there are 19 days that were labeled as Semi-Normal days in the test dataset, but are misclassified as Abnormal days. These days form a set, which is named as MisDays (Semi/Abnormal) and are listed below.

MisDays (Semi/Abnormal):

"2019-02-18, 2019-02-21, 2019-04-01, 2019-04-20, 2019-06-17, 2019-06-18, 2019-06-19, 2019-06-21, 2019-06-24, 2019-06-25, 2019-06-30, 2019-07-12, 2019-07-20, 2019-07-22, 2019-07-23, 2019-08-09, 2019-08-19, 2019-08-21, and 2019-12-03".

In addition to MisDays (Semi/Abnormal), 3 days of Abnormal days in train dataset are misclassified as Semi-Normal days. Similarly, these days are named MisDays (Abnormal/Semi) and listed below.

MisDays (Abnormal/Semi):

"2019-02-20, 3019-08-20, and 2019-12-01".

4.2.1.1. Misclassified days: Equity index

Figure 26 shows the equity index scatter plot of all airlines on these misclassified days. The diagonal line represents a situation, at which both Noon and Midnight equity indices are equal. As it is seen, points are scattered around the diagonal line on MisDays (Semi/Abnormal). On the other hand, on MisDays (Abnormal/Semi), points are scattered above and below the diagonal line randomly without showing any particular pattern.

Visually speaking, the results denote that there is no significant relationship between misclassified days and the difference of equity induces of noon and midnight. Thus, the hypothesis of experiencing high values of equity index at noon but low values of that at midnight does affect Semi-Normal days to be misclassified as Ab-normal days is refused.

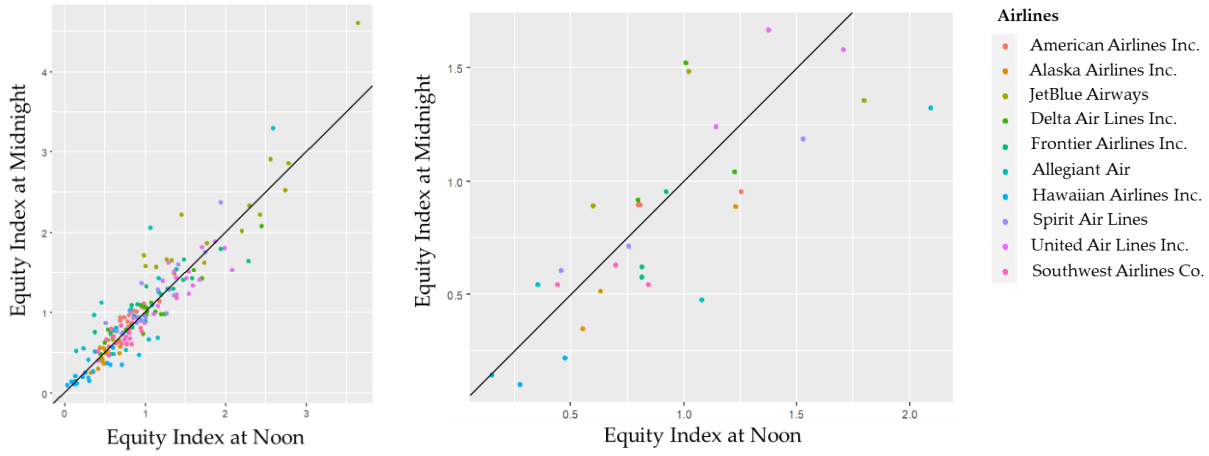


Figure 26. Equity index scatter plot for misclassified days: (a) MisDays (Semi/Abnormal), and (b) MisDays (Abnormal/Semi).

4.2.1.2 Misclassified days: Cancelation rate analytics

To analyze the misclassified days, particularly MisDays (Semi/Abnormal), delay profiles of days in the test dataset were created and shown in Figure 27. In this figure, actual Abnormal days and MisDays (Semi/Abnormal) are highlighted for a better comparison. As it is seen, MisDays (Semi/Abnormal) have similar pattern with Abnormal days up to almost 4:00 p.m., which cannot be trained by the methods. It is seen that the slope of Abnormal days is getting steeper than MisDays (Semi/Abnormal) after 4:00 p.m. It is hypothesized that (difference in steepness of slopes) happens as the results of more canceled flights on MisDays (Semi/Abnormal).

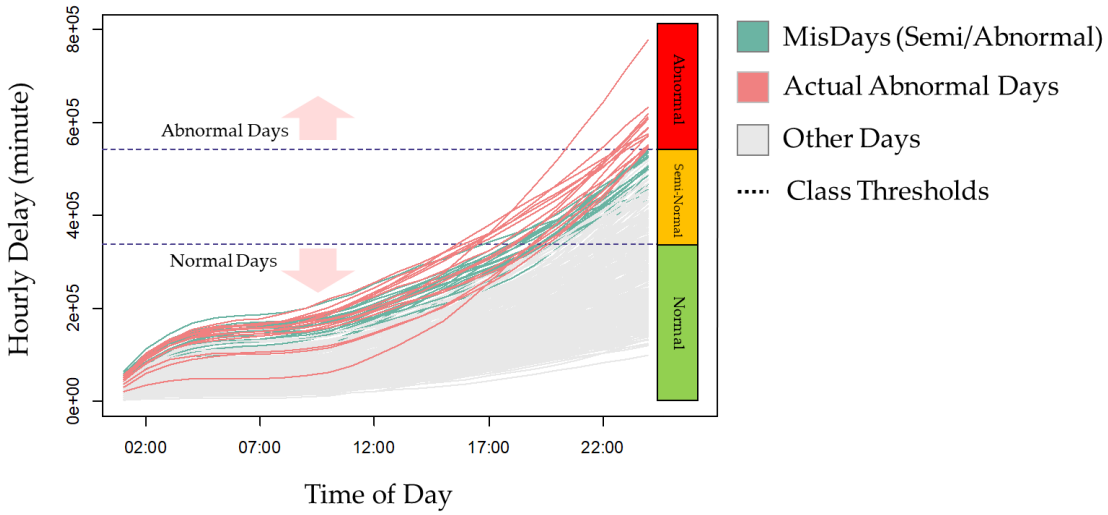


Figure 27. Delay profiles for MisDays (Semi/Abnormal) analysis.

Cancelation rate of MisDays (Semi/Abnormal), and other days is shown in Figure 28. Average cancelation rate of MisDays (Semi/Abnormal) is 2.52 percent ($\mu_{\text{MisDays (Semi/Abnormal)}} = 2.52\%$), while that value is 1.91 percent ($\mu_{\text{Other}} = 1.91\%$), for other days. To investigate if these values are statically different, a null and an alternative hypothesis were formed as follow:

Null (H_0): $\mu_{\text{MisDays (Semi/Abnormal)}} = \mu_{\text{Other}}$
Alternative (H_1): $\mu_{\text{MisDays (Semi/Abnormal)}} \neq \mu_{\text{Other}}$
Critical region: Reject H_0 if $|T| > 1.96$ (95% confidence level)

Obtained results indicate the value of 1.47 for $|T|$. Meaning that, we failed to reject the null hypothesis. Statically speaking, there is no difference between cancelation rate of MisDays (Semi/Abnormal) and other days. It should be noted that the estimated average cancelation rate of MisDays (Abnormal/Semi) is 6.15 percent. However, due to the very small number of MisDays (Abnormal/Semi), it is not reasonable to compare average cancelation ratio values of MisDays (Abnormal/Semi) and other days.

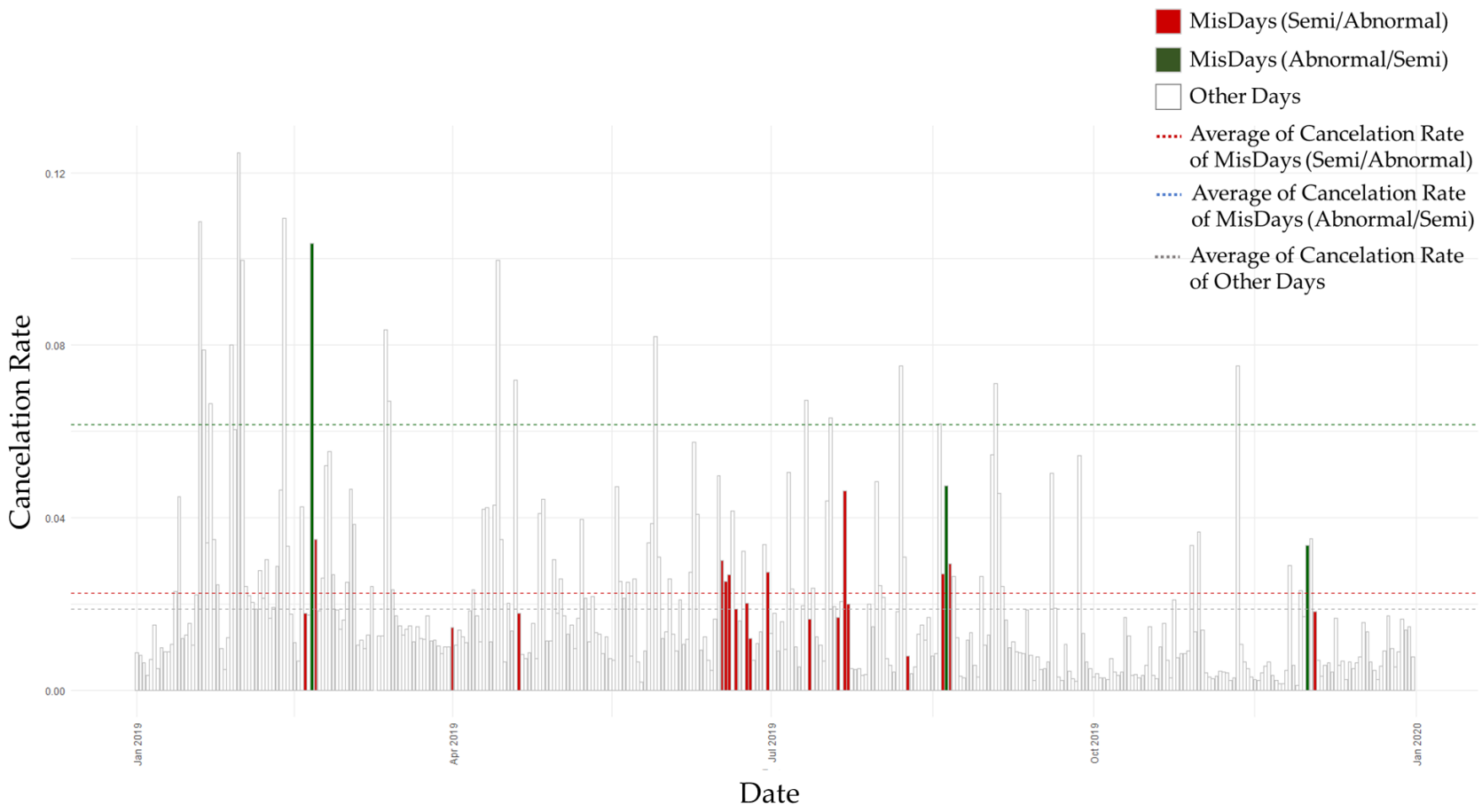


Figure 28. Cancelation rate profile.

To further analyze MisDays (Abnormal/Semi), these days and actual Semi-Normal days are highlighted in delay profiles shown in Figure 29. MisDays (Abnormal/Semi) are, in overall, having delay as equal to an average delay of actual Semi-Normal days. This trend continues until 5:00 p.m., and MisDays (Abnormal/Semi) curves start to diverge from actual Semi-Normal Days after that time. It can be concluded that these abrupt changes in MisDays (Abnormal/Semi) led to misclassifying of these days. In other words, since these changes happened in the future of the modeling time, the models were not capable of predicting classes correctly.

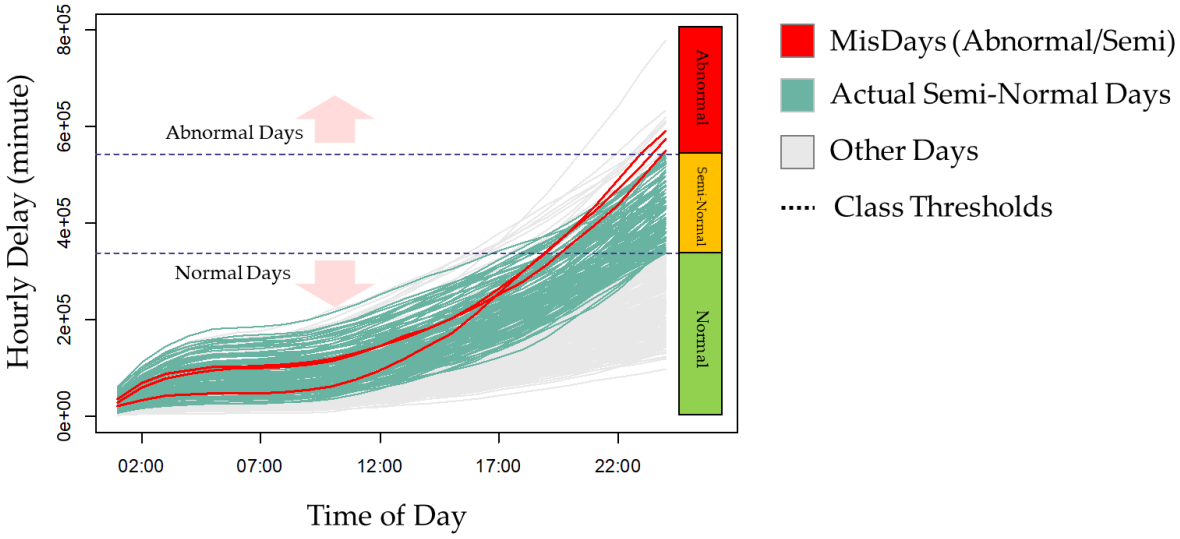


Figure 29. Delay profiles for MisDays (Abnormal/Semi) analysis.

4.2.2. Adding cancelation rates

Investigating cancelation rates, it was decided to develop new models by adding a cancelation-related variable to previous models and check the performance of the new models. With respect to the fact that prediction models are utilizing information up to noon (maybe 1:00 p.m. or 2:00 p.m.) of a given day and assuming of having information of canceled flights up to six hours ahead of time, we added a variable indicating cancelation rate of a given day up to 6:00 p.m. of that day. The results of new models are reported in Table 10 and Table 11. New models are tagged with CR (Cancelation Rate). Table 10 and Table 11 compare new models' results with results that were obtained previously. Since LSTM and RF did not perform as well as SVM, it was decided not to list the results of new models of these methods. Results indicate that new models are more accurate in terms of overall accuracy rate. Additionally, the detection rate of Semi-Normal days substantially increased from approximately 70 percent to approximately 80

percent. However, the detection rate of Abnormal days has decreased significantly. No significant change was observed in detection rate of Normal days.

Table 10. Confusion matrix for delay severity prediction models with different sizes of history response and added variable.

Delay Severity Level	Prediction Method*	Classified p_{ij}			r_{ij}		
		Normal	Semi-Normal	Abnormal	Normal	Semi-Normal	Abnormal
Normal	SVM (12)	188	46	0	**80.34%**	19.66%	00.00%
N_1= 234	SVM (13)	195	39	0	**83.33%**	16.67%	00.00%
	SVM (14)	201	33	0	**85.90%**	14.10%	00.00%
	SVM (15)	206	28	0	**88.03%**	11.97%	00.00%
	SVM (12 + CR[a])	192	41	1	**82.05%**	17.52%	00.43%
	SVM (13 + CR)	193	40	1	**82.48%**	17.09%	00.43%
	SVM (14 + CR)	194	39	1	**82.91%**	16.67%	00.43%
	SVM (15 + CR)	199	34	1	**85.04%**	14.53%	00.43%
Semi-Normal	SVM (12)	15	82	19	12.93%	**70.69 %**	16.38%
N_2= 116	SVM (13)	14	84	18	12.07%	**72.41%**	15.52%
	SVM (14)	15	82	19	12.93%	**70.69%**	16.38%
	SVM (15)	9	87	20	7.76%	**75.00%**	17.24%
	SVM (12 + CR)	12	93	11	10.34%	**80.17%**	9.48%
	SVM (13 + CR)	15	92	9	12.93%	**79.31%**	7.76%
	SVM (14 + CR)	13	95	8	11.21%	**81.90%**	6.89%
	SVM (15 + CR)	9	97	11	7.76%	**83.62%**	9.48%
Abnormal	SVM (12)	0	3	11	00.00%	21.43%	**78.57%**
N_3= 14	SVM (13)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (14)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (15)	0	3	11	00.00%	21.43%	**78.57%**
	SVM (12 + CR)	1	5	8	7.14%	35.71%	**57.14%**
	SVM (13 + CR)	1	5	8	7.14%	35.71%	**57.14%**
	SVM (14 + CR)	0	4	10	00.00%	28.57%	**71.43%**
	SVM (15 + CR)	0	3	11	00.00%	21.43%	**78.57%**

* Numbers in parentheses account for the size of response history.

[a] CR indicates that the model used the cancelation rate variable.

All in all, adding cancelation ratio to the models causes more Abnormal days to be predicted as Semi-Normal days. Plus, it leads to less Semi-Normal days to be predicted as Abnormal days. It appears that cancelation ratio solved the issue of MisDays (Semi/Abnormal) due to difference in the steepness of slopes (Figure 27). Meanwhile, it over-informs the models, which causes misclassification.

Table 11. Prediction accuracy rates of new and previous SVM models.

Size of History Response	Overall Accuracy Rate	
	Prediction Method	
	SVM	SVM + CR[a]
12	77.20%	80.49%
13	79.67%	80.49%
14	80.77%	82.14%
15	83.52%	84.34%

[a] Model used the cancelation rate variable.

Figure 30 shows delay profiles of each classified level based on the results achieved from the most accurate model, SVM (15 + CR). Comparing Figure 30 and Figure 21, it is seen that misclassified days are relatively decreased in the former one.

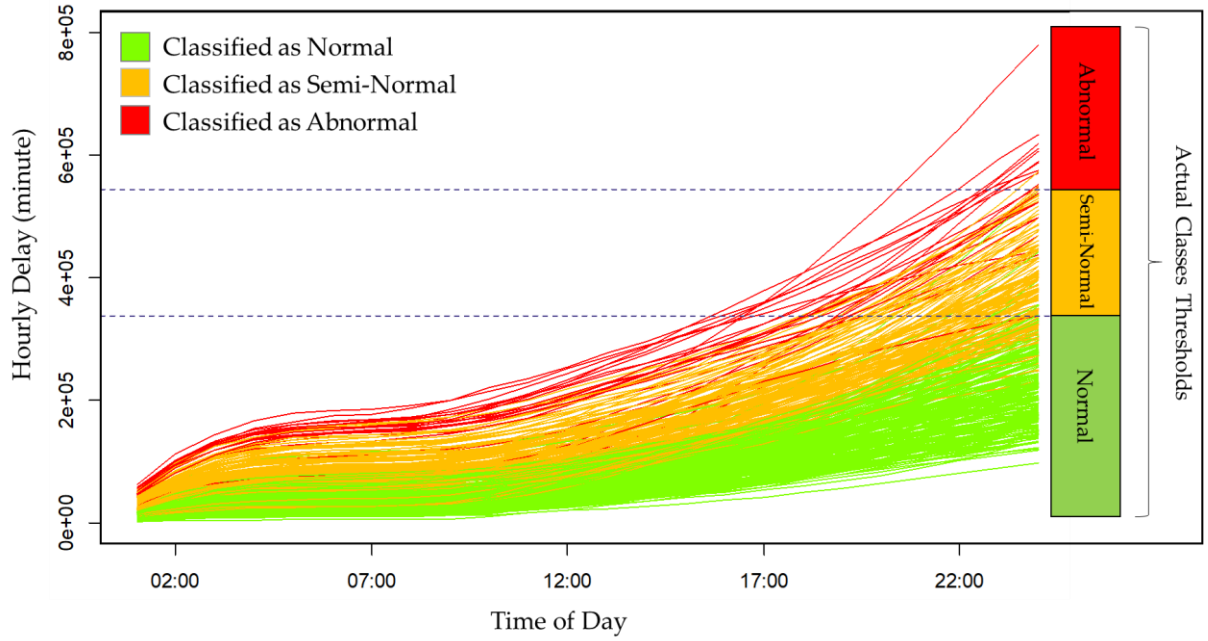


Figure 30. Classified delays based on SVM (15 +CR) vs actual delay profiles.

5. Discussion and Conclusion

The goal of this study has been to predict the cumulative arrival flight delay based on different methods employed and help solve a major problem for airlines that costs them millions. Beginning with clean-up and analysis, the two-year aircraft delay dataset was reduced to a size appropriate for the study, with relevant data only and by removing outliers. This also included data manipulation of time zones and adjustment of other attributes to make the predictive modeling feasible.

In the initial analysis the Loess Regression model was applied to see when the minimum and maximum delay happens, which the findings showed the minimum hourly delays happened at 6 am, and the maximum delays occurred towards the end of the day, which was expected. Visualizations were to show each of these models and detailed delay visualizations. A temporal heat map also showed that flights arriving after 8pm on Thursdays and Fridays are subject to higher delays.

Both regression and classification methods were employed in this study. In the regression approach, the time series models were applied, and their performance were tested. A recursive regression method (ATASH) was also proposed to address time series modeling. The performance of this method was acceptable and comparable to a well-known advanced machine learning method, Long Short-Term Memory (LSTM). Between the two approaches taken by the team with regards to time series regression models, the LSTM model was found to be better than that of ATASH, as a result ATASH was not used for time series regression for different airlines. Besides the LSTM, two other classification methods of Support Vector Machine (SVM) and Random Forest (RF) were applied. Comparing the accuracy and the detection rate resulted of these methods, SVM had the performance.

Equity index plots were developed for some of the major carriers, those with the highest market share. No pattern was observed with the class of delay in these plots. The conclusion drawn from the analysis is that there's almost no correlation between equity indices of airlines and classes of delay. Findings also included that some of the smaller airlines like JetBlue and Frontier had some of the greater contribution to delays, as compared to others with greater market share.

There were some days that were misclassified. An analysis conducted to see if there was any relationship between misclassified days and the equity indices. It was seen that there was no significant relationship denoted. A further analysis of cancellation rate and misclassified days (Semi normal and Abnormal) was conducted, and results found there

was difference in cancelation rate of misclassed and other days, and a cancellation rate of 6.15 percent was found for the misclassified days.

In further analysis of cancellation rates, there was a cancelation related variable added to the previous models mentioned above. During the course of this analysis, it was found that RF and LSTM did not perform as well as SVM, so the results of SVM were presented instead. The SVM models with the classification rate indicator are more accurate than previous results. Another implication of the analysis was that although the detection rate of semi-normal days increased substantially, the detection rate of abnormal days decreased significantly. However, overall misclassified days were decreased significantly.

As seen within this report, regression and classification models can be applied to predicting flight delays. Regression and classification machine modeling techniques are considered to be supervised learning models. Individuals are not limited to applying just Long-Short Term Memory and Support Vector Regression. It would be interesting to see how other models such as linear regression, logistic regression, and decision trees can be applied to the flight delays to make a prediction of delay outlook for the given day. Of course, there can be different parameters that can be applied to these models that can potentially help to provide better accuracy. As well as the addition of potentially other important variables that can help to determine the flight delay such as weather, maintenance, cargo, and bird strike.

Acknowledgment

This research has been conducted as DAEN 690: Data Analytics Engineering MS Capstone Project under sponsorship of Center for Air Transportation Systems Research (CATSR). The authors would like to thank CATSR Director Dr. Lance Sherry and Ph.D. student Sasha Donnelly for their advice, counsel, and expertise about the aviation systems. Special thanks also go to DAEN Capstone Project Coordinator F. Brett Berlin for his support and assistance with this project.

An essential element of this study was accession to data sources and information about flights across the country. The authors wish to thank BTS for collecting and archiving the airline on-time performance data.

References

Airline On-Time Performance Data. Bureau of Transportation Statistics. n.d. https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time (accessed May 2, 2020).

Breiman, Leo. "Random Forests." *Machine Learning* 45 (2001): 5-32.

Chakrabarty, Navoneel. "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines." *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 2019: 102-107.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357.

Greff, Klaus, Rupesh Srivastava, Jon Koutn´ık, Bas Steunebrink, and Jurgen Schmidhuber. "LSTM: A Search Space Odyssey." *IEEE Transactions on Neural Networks and Learning Systems* 28, no. 10 (2017): 2222-2232.

Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques.* Elsevier, 2011.

Khanmohammadi, Sina, Salih Tutun, and Yunus Kucuk. "A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport." *Procedia Computer Sceince* 95 (2016): 237-244.

Phi, Michael. *Illustrated Guide to LSTM's and GRU's: A step by step explanation.* Towards Data Science. Sep 24, 2018. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (accessed May 2, 2020).

Sherry, Lance. "Air Transportation Delay Analysis Workbook." 2010.

Stehman, Stephen V. "Selecting and interpreting measures of thematic classification accuracy." *Remote Sensing of Environment* 62, no. 1 (1991): 77-89.

Walimbe, Rohit. *Handling imbalanced dataset in supervised learning using family of SMOTE algorithm.* Data Science Central. April 24, 2017. https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family (accessed April 25, 2020).