



# US flights Machine learning Project



Sean Chang  
[sean.chang@duke.edu](mailto:sean.chang@duke.edu)



# Outline

---

- Goals
- Datasets
- Exploratory data analysis
- Flight delays modelling
- Conclusions

# Goals

---

# Goals and Impacts

---

- Depict global pictures of US flight transportations in recent years.
- Develop an accurate flight delays prediction system based on transportation and weather data.
- Save billion of dollars every years of flights delays due to additional hotel/ taxi/ tickets expenses [1].
- Provide more than 800 million per year US air travelers a reliable timeline when scheduling flights & travels [2].

# Datasets

---

# Bureau of Transportation Flights Statistics

US Flights data since 2005 include:

- Flight schedule date.
- Carrier.
- Flight/ tail numbers.
- Flight origin/ destination.
- Actual departure/ arrival time.
- Cancellation code.
- Cause of Delay (5 categories).

The screenshot displays the Bureau of Transportation Statistics website. At the top, the header includes the United States Department of Transportation logo and the text "OFFICE OF THE ASSISTANT SECRETARY FOR RESEARCH AND TECHNOLOGY". Below this, the "Bureau of Transportation Statistics" title is prominent. A navigation bar contains links for "About BTS", "BTS Press Room", "Data and Statistics", "Publications", and "Search". The "Data and Statistics" link is selected, leading to the "OST-R > BTS" page.

The main content area features the "TranStats" logo and a search bar with the text "Search this site:". To the right of the search bar is a "Go" button. Below the search bar is a link for "Advanced Search".

On the left side, there are two sections: "Resources" and "Data Tools". The "Resources" section includes links for "Database Directory", "Glossary", "Upcoming Releases", and "Data Release History". The "Data Tools" section includes links for "Analysis", "Table Profile", and "Table Contents".

The main search results area is titled "On-Time : On-Time Performance". It includes a "Download Instructions" link and a "Latest Available Data: December 2015" label. There are filters for "Filter Geography" (set to "All") and "Filter Year" (set to "2015"). Below these filters are checkboxes for "Prezipped File", "% Missing", "Documentation", and "Terms".

The search results are displayed in a table with two columns: "Field Name" and "Description". The table includes the following fields:

Field Name	Description
<input type="checkbox"/> Year	Year
<input type="checkbox"/> Quarter	Quarter (1-4)
<input type="checkbox"/> Month	Month
<input type="checkbox"/> DayofMonth	Day of Month
<input type="checkbox"/> DayOfWeek	Day of Week
<input type="checkbox"/> FlightDate	Flight Date (yyyymmdd)
<input type="checkbox"/> UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.

# ASOS-AWOS-METAR Data

Detailed hourly weather data of US airports:

- Air Temperature
- Dew Point Temperature
- Humidity
- Wind speed
- One hour precipitation
- Pressure altimeter
- Visibility
- Wind gust

Select Network

1) Select Station/Network by clicking on location:

Select Widget for IA\_ASOS Network

Sort Available Stations: ▼

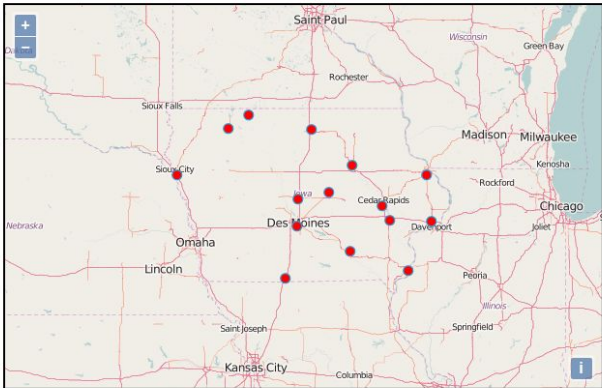
[ALO] Waterloo  
[AMW] Ames  
[BRL] BURLINGTON  
[CID] CEDAR RAPIDS  
[DSM] Des Moines

Enter some text here to filter

+ Add Selected Add All

Selected Stations:

Remove Selected Remove All



Download Variable Description

station: three or four character site identifier

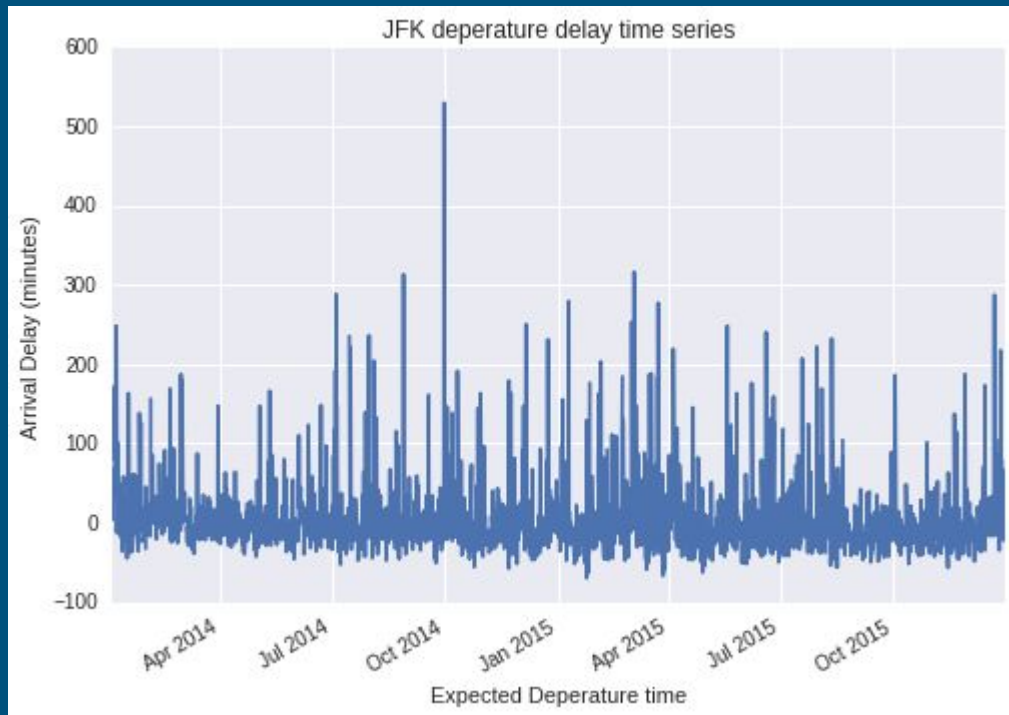
# Exploratory Data Analysis

---

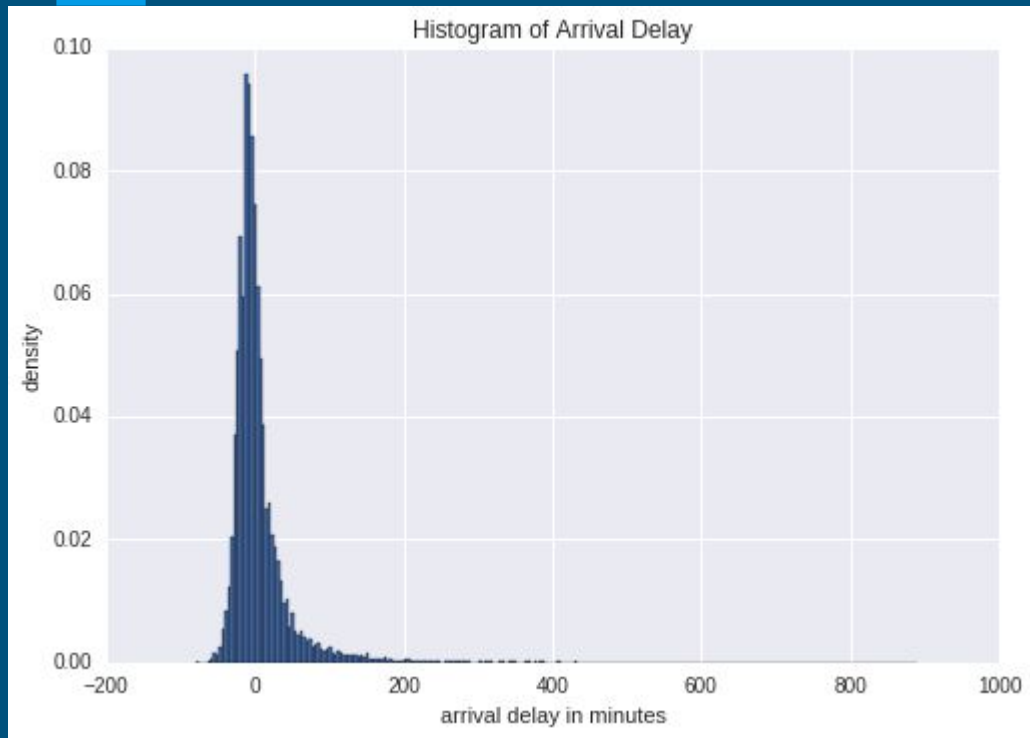


# Flight delays at JFK, 2014-2015

- Delay time series includes all flights in JFK in 2014-2015.
- Hardly to find any periodicity of this series, in other words, overall temporal trend of delays seems to be quite weak.
- One goal of this project is modelling the 'outliers', where delays more than 200 minutes.



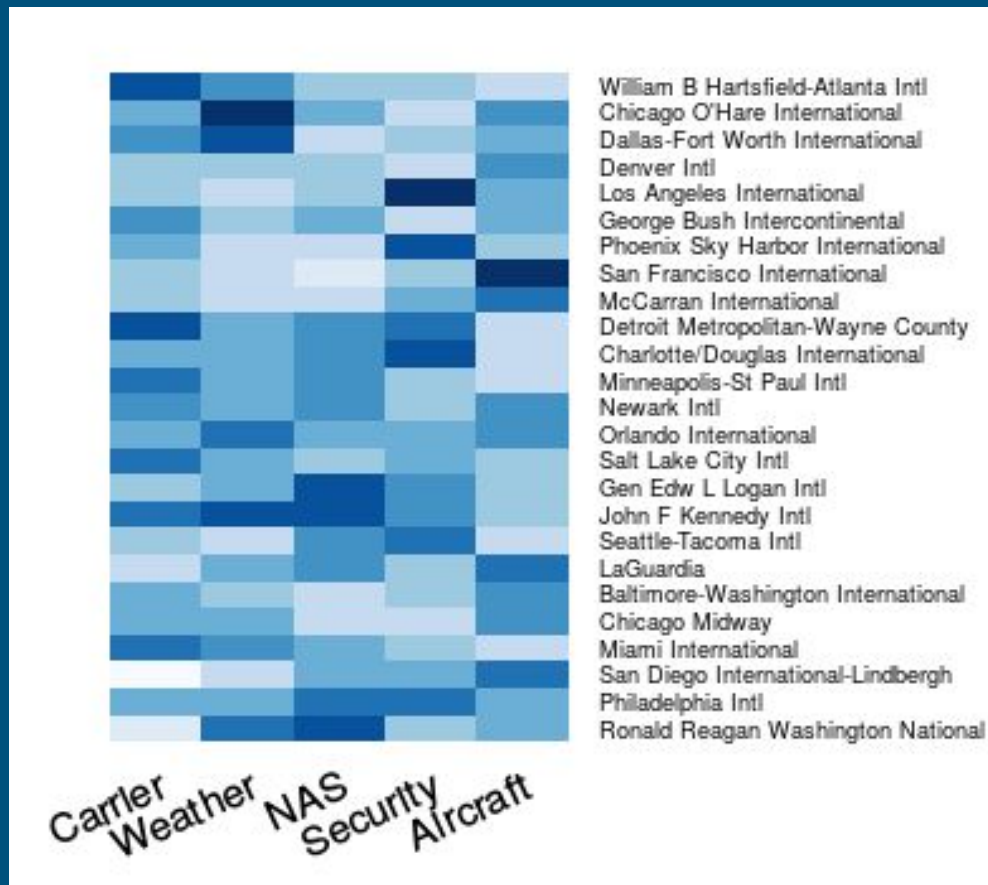
# Flights delay distribution



- Actually nearly 60% flights arrive before the schedule, among these early arrival flights, their average is 15 mins earlier.
- However, delays distribution is skewed: if delayed, on average 35 minutes is expected, and it can be as worse as hours (90% percentile).
- Maybe this is why many of us remember these awful delays instead of on-time experiences.

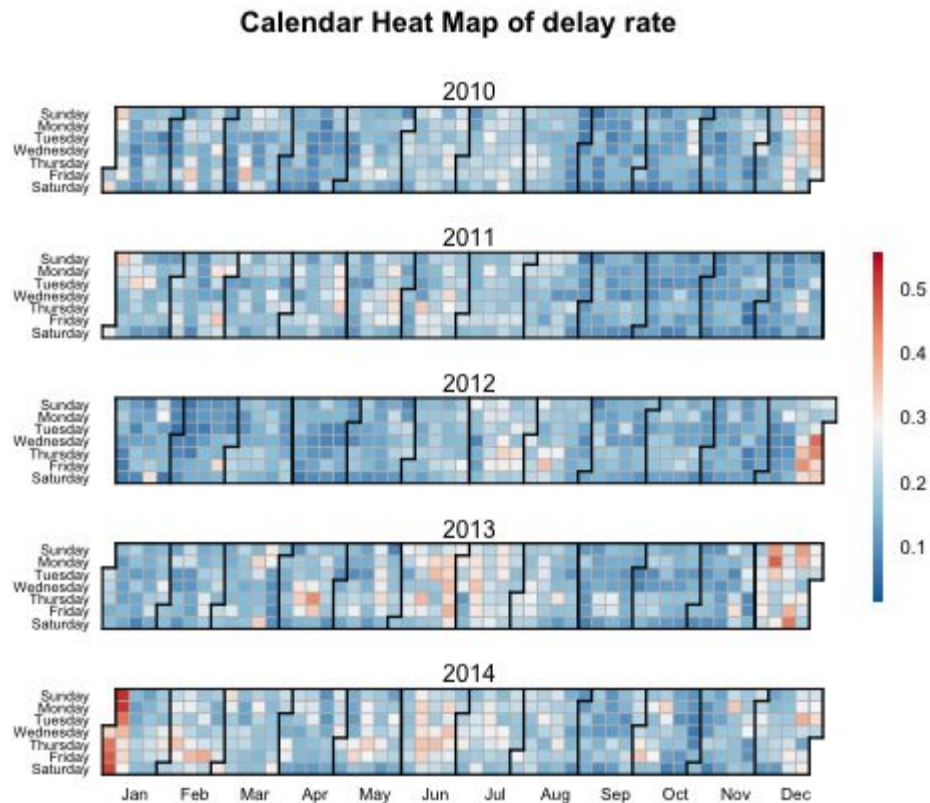
# Heatmap

- Weather delays are much more common in Chicago and New York (JFK) than in Phoenix and San Francisco.
- Security is a major reason for delay in LAX but not in Denver.
- San Diego and Ronald Reagan Washington airports may have good carrier control management, resulting in less carrier delays.

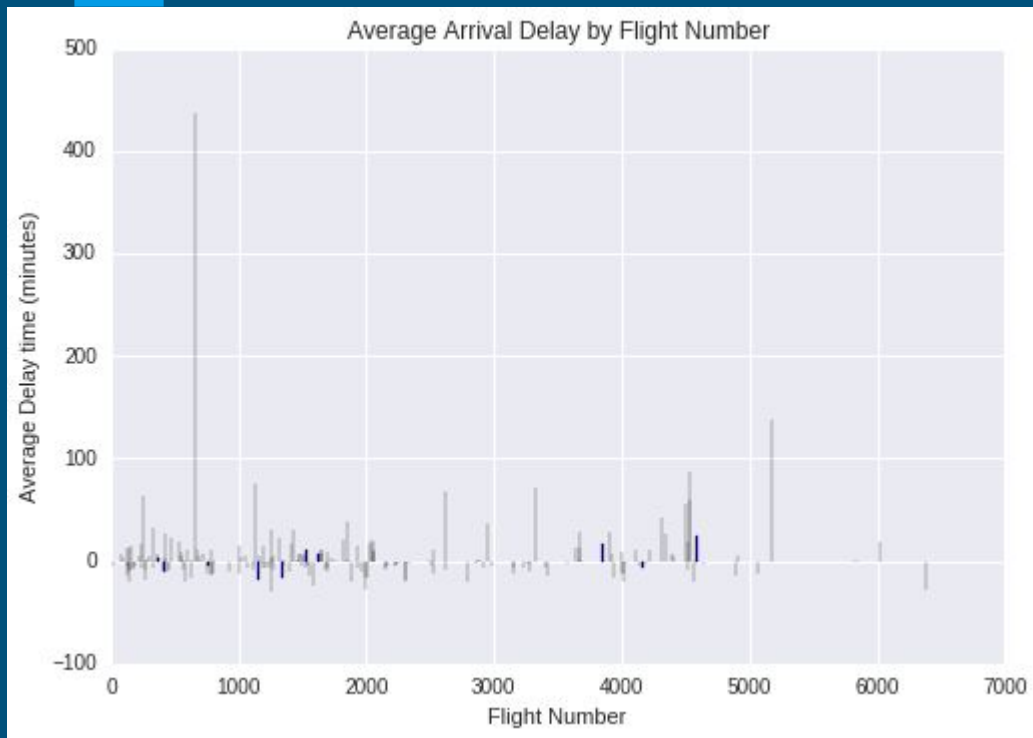


# Calendar Plot of delay rates

- On each date,  $(\# \text{ delays}) / (\# \text{ flights})$  is reported and colored.
- Higher rates in Jan, July and Aug, match the traveling seasons. Sometimes delays are overwhelming, e.g. in the first week of 2014, more than 50% flights were delayed.
- From 2012-2014, the rate goes up gradually. The causes of this needs to be verified in further studies.



# Delays aggregated by Flight Number



- Flight number specifies a particular airline and route.
- Flight number is a good indicator for flight delays: some flights keep good on-time records, but some are notoriously bad.
- This feature will be used in advanced models next section.

# Flight Delays Modelling

---

# Methodologies

---

In order to build and test prototypes quickly, consider:

- Flights between 5 large US airports: Atlanta (ATL), Washington (IAH), New York (JFK), Los Angeles (LAX) and Chicago (ORD).
- Random sampling 10% of above data, use 2014 data as training data, 2015 data as testing data.
- Make predictions and access models based on  
(a) root mean square error (MAE), and (b) average absolute error (RMSE).

# Naïve Approaches

---

## Average of previous delays

- Report average delays on the same route.
- And same flight number?
- Same time? Same weather condition?
- Obviously, nearly impossible to find exact data points having the same conditions. This approach is limited.

## Linear regression

- Regress flight delay on origin + dest airports + Flight Number + other features.
- However, tons of categorical variables makes regression computationally expensive.
- Easily overfitting.

Both models have MAE ~ 25 minutes and RMSE ~40 minutes.



# Random forest approach

---

## Random Forests regressions.

- Features include route, carrier, flight number and weather.
- Weather of both departure/arrival airports closest to the scheduled departure time is considered.

## Other benefits:

- Not using all the features each time making predictions to speed up the training.
- Prevent overfitting by averaging many decision trees.
- Reduce biases by selecting a few variables each time.

# Results 1: Important features

---

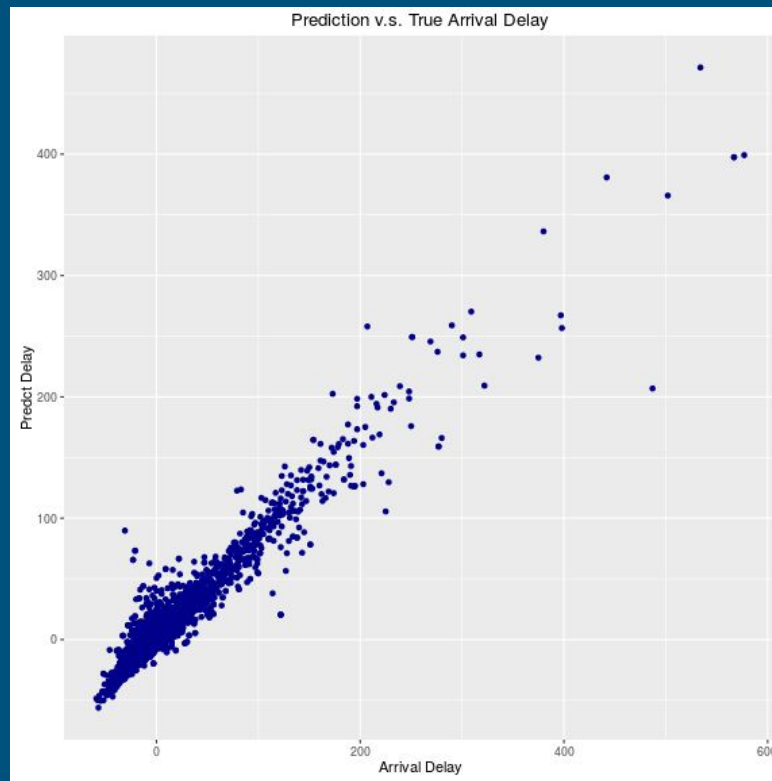
Importance is measuring by the increase in mean-square-error of predictions (estimated with out-of-bag-CV) as a result of variable.

Most important features are:

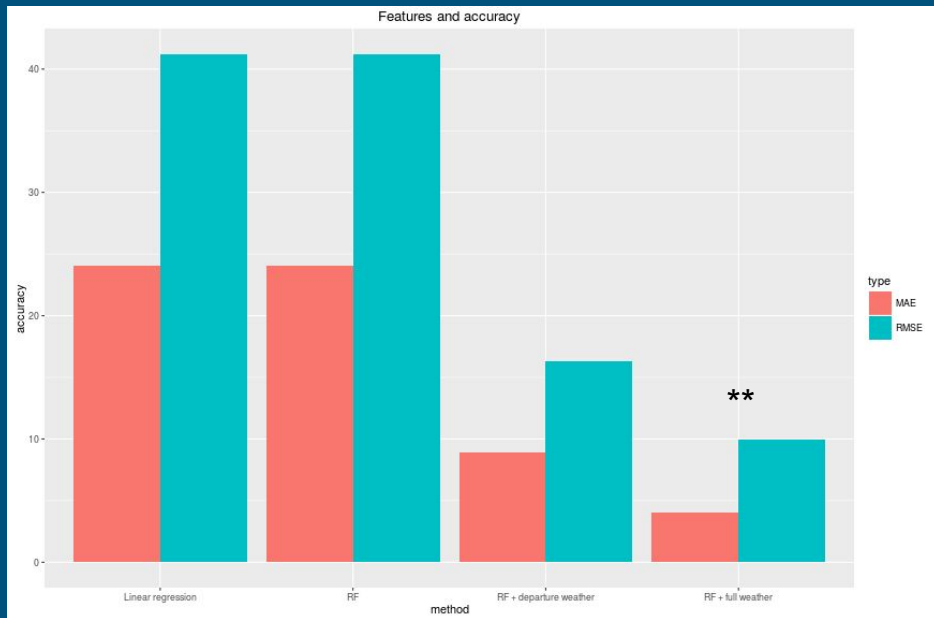
- Destination wind speed
- Destination dew point temperature
- Destination Pressure altimeter
- Carrier
- Origin Pressure altimeter
- Flight Number

# Results 2: Predictions

- Scatter plot of arrival delays vs predicted delays of the testing data.
- The random forest model forecasts early arrivals and short delays really well (prediction vs truth are on the  $x=y$  line).
- For long delays (>3 hours, less than 1% of data), the model is still able to make predictions with error ~30 minutes, achieving such accuracy for outliers is remarkable.



# Results 3: Comparisons



- The chart: predictions mean absolute error (red) and root mean square error (teal) of four models:
  - Linear regression
  - random forest (no weather)
  - random forest (+departure weather)
  - \*\*random forest (+dept & destination weather)
- Weather data is extremely important for modelling flight delays. Without weather info, predictions can only reach 40 mins RMSE accuracy.
- Random forest full model has MAE 6 minutes and RMSE 10 minutes, >65% improvements from the naive approaches.

# Conclusions

---

# Summary and future work

---

- Mining useful information from massive flight and weather datasets, and demonstrate data visualization findings comprehensively.
- The random forest model predicts flight delays on average having less than 10 minutes , improving from 40+ minutes of empirical approaches.
- Similar approach can be used to model the cancellation rates (<1% of flights).
- The model can incorporate time series to capture temporal delays caused by traffic jams or special events such as Super Bowl.
- Modelling interactions between airports potentially will provide more instantaneous delay information.

# Thank you



Any feedback would be greatly appreciated!

# Appendix

---

- AWOS <https://mesonet.agron.iastate.edu/request/download.phtml>
- [Bureau of Transportation Flights Statistics.](#)
- [1] all, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A. A., and Zou, B. (2010), "Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States," .
- [2] Total Passengers on U.S Airlines and Foreign Airlines U.S. Flights Increased 1.3% in 2012 from 2011, [http://www.rita.dot.gov/bts/press\\_releases/bts016\\_13](http://www.rita.dot.gov/bts/press_releases/bts016_13)