Building an infrastructure for investigating the effects of weather conditions on flight delays

Final Report

Vera Lo
Industrial and Operations Engineering Department
University of Michigan

Faculty Advisor: Professor Amy Cohn

September 2013

Table of Contents

I. Introduction

1. U.S. domestic airline industry

The U.S. domestic airline industry is an economically-significant industry with an abundant amount of capital invested in it. Carrying over one-third of the world's total air traffic and over 737 million of passengers in 2011, the industry was valued at USD$187 billion in 2011 with a forecast value of USD$316 billion in 2016, undertaking an estimated increase of 70%.[1] The industry not only plays a critical role in the U.S. economy by contributing to 5.2% of the nation's GDP, but also is important in building connectivity in the global economy.[2]

2. Flight Delays

Given the complexity and the large extent of the interdependencies between airports, aircraft, passengers, airlines, control centers, etc. of the national aviation system, flight delays occur frequently. Based on data collection in the past ten years from 2003 to 2012, more than 688,000 flights are delayed among the 3 million flight operations each year, accounting to approximately 20% of the total flight operations.[3]

The costs and impacts of flight delays are further augmented in a network structure. The inter-connectivity within the system creates a "ripple" effect as delays originating at an airport propagate forward to multiple airports during the course of a single operational day. According to a research commissioned by the Federal Aviation Administration, the U.S. economy incurred a cost of USD$32.9 billion due to flight delays in 2007. Half the cost was borne by passengers.[4]

3. Weather Delays – the main cause of flight delays

Weather delays are extremely significant among the different causes of delays. Severe weather conditions, such as thunderstorms, turbulences, in-flight icing, low ceiling and visibility level etc., often interrupt operations at local airports, resulting in reduction in airport capacities, hindrance in ground operations, or even closure of airports in some extreme cases. Hence flight operations are affected; as a result, delays and re-routing of flights create additional costs to the flight company due to loss revenue, excess maintenance costs, and additional fuel costs.

---

[1] MarketLine. (2012, October 31). United States – Airlines. [industry profile]. Retrieved from MarketLine Advantage database.

[2] U.S. Department of Transportation, Federal Aviation Administration. (2011). *The economic impact of civil aviation on the U.S. economy*. Retrieved from ATO Communications website: http://www.faa.gov/air_traffic/publications/media/FAA_Economic_Impact_Rpt_2011.pdf

[3] *On-time performance - flight delays at a glance*. (2013). Retrieved from http://www.transtats.bts.gov/HomeDrillChart.asp?URL_SelectYear=2013&URL_SelectMonth=6&URL_Time=1&URL_Selection=1

[4] Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States, NEXTOR (2010, December 16). Retrieved from http://www.isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

4. Goals and Objectives

Realizing the importance and value of mitigating weather delays, we collaborated with Southwest Airlines to conduct a weather index project. The primary goal for this summer research project is to develop a sophisticated and dynamic database tool that contains ten years' worth of weather data and the corresponding flight data. The database aims to allow users to conduct quick queries, as well as to expand the database and add new data to it easily as time progresses. Data stored spans all major airports and carriers in the United States, and enables a wide range of detailed analyses linking weather at origin/destination airports and flight delays.

Potential questions that could be addressed with future analyses, include:

- Which weather factor impacts on-time performance the most?
- How long do weather delays typically last?
- Weather at which airport has the biggest impact on the system as a whole?
- Which predicts on-time performance better:
    - the weather at the origin airport at the actual time of departure,
    - the weather at the arrival airport at the scheduled time of arrival,
    - the weather at the arrival airport at the actual time of arrival, or
    - the weather at the arrival airport at the time of departure, etc.?
- Under which weather conditions does Southwest Airlines perform better than other carriers? Under which weather conditions is the on-time performance of Southwest Airlines worse?

Ultimately, we do not simply seek to understand flight delays, but rather use this understanding to reduce the impact of delays.

II. Data Processing

1. Weather Data

The source of the weather data used is the Integrated Surface Data (ISD) generated by the National Climate Data Center to collect hourly and synoptic weather observations. This set of data is accessible online via NCDC's Climate Data Online system, FTP, and GIS services.[5] Seven weather factors are included in the database, namely wind direction, wind speed, ceiling height, visibility distance, air temperature, dew point, and sea level pressure. The database currently holds 41.6 million of hourly weather records measured at weather stations of 316 major U.S. airports.

2. Flight Data

We used the Airline On-time Statistics on the TranStats website of the Bureau of Transportation Statistics as our source of flight data. From the website, we were able to extract information regarding the time period, airline, origin, destination, departure on-time performance, arrival on-time performance, and flight summaries of all domestic flights for each month since 1987. Each year, there are more than 67,000 domestic flight operations across the country with an average of 18,000 flights per day. For this project, we collected data for the past 10 years from 2003 to 2012, which aggregates to 68 million of records.

3. Obstacles and Challenges

Although the data used in this project is readily available on the internet, processing the data into usable formats for analysis posed many major obstacles that took months to overcome, such as:

i. Handling large data –
   Downloading and storing all the relevant data from the internet manually is an extremely time-consuming task. In addition, analyzing all ten years of data simultaneously far exceeds the capability of Microsoft applications or any other commonly used software.

ii. Converting time zones and observing daylight savings time –
   Processing the flight data is especially challenging in this aspect, as there are four different times listed in each record (scheduled and actual times for both departure and arrival) which could be from different time zones and on different days when the flight is scheduled around midnight. Additionally, not all U.S. states/cities observe Daylight Savings Time.

[5] Burroughs, J. (2008, August 20). *Integrated Surface Database*. Retrieved from http://www.ncdc.noaa.gov/oa/climate/isd/

iii. Interpreting units –
Each weather data file was coded in ASCII character format and contains much supplementary information, such that the raw data needs to be processed before they can be used for importing into the database and conducting statistical analyses. For the convenience of future domestic airline users, all weather data in metric units were converted to conventional units in the U.S. aviation industry and were recalibrated based on the given scaling factor.

iv. Linking the flight and weather data –
The two sets of data can only be linked given a specific date, time, and airport combination, but there are limited ways to do so, especially taking the aforementioned constraints in account as well.

4. Solutions

In face of the challenges mentioned above, we considered a number of options and experimented to address the issues, including:

i. Manually downloading and parsing the files with the use of self-written Visual Studio codes
ii. Using self-developed Macro-enabled Excel files to prepare the files for merging
iii. Storing the data on a file server, and performing analysis on Excel using the VLookup function
iv. Setting up a database for Microsoft Access, and conduct analysis using SQL query statements

Ultimately, due to the voluminous amount of data involved, we wrote Python scripts to automate the downloading and uploading processes of all data files. We also requested university support to host a large-scale database using phpMyAdmin, an open source tool intended to handle the administration of MySQL via a web browser. It not only provides convenient access to users as a web-based tool, but also enables users to execute SQL query statements and manage databases, tables, fields and rows. On the whole, the software suits our needs well.

III. Analysis

As mentioned in *Section I Introduction*, the database serves as a platform to conduct a wide range of analyses using SQL queries. With the all-encompassing data in the database, we formulated several ways of analyzing the data by selecting different parameters to extrapolate interesting patterns that enable us to better understand flight delay. In this section, I will discuss some examples of the analyses conducted.

1. Analyses using weather data solely

The database allows users to easily tap across all ten years' of weather data to investigate the weather characteristics at each major airport. For example, low-lying clouds and dense fog are known to cause a large extent of flight delays at the San Francisco International Airport (SFO). To explore this phenomenon further, we plot the hourly trend of visibility distance at SFO, as shown in Figure 1 below. Anomalies can also be identified easily with these charts.

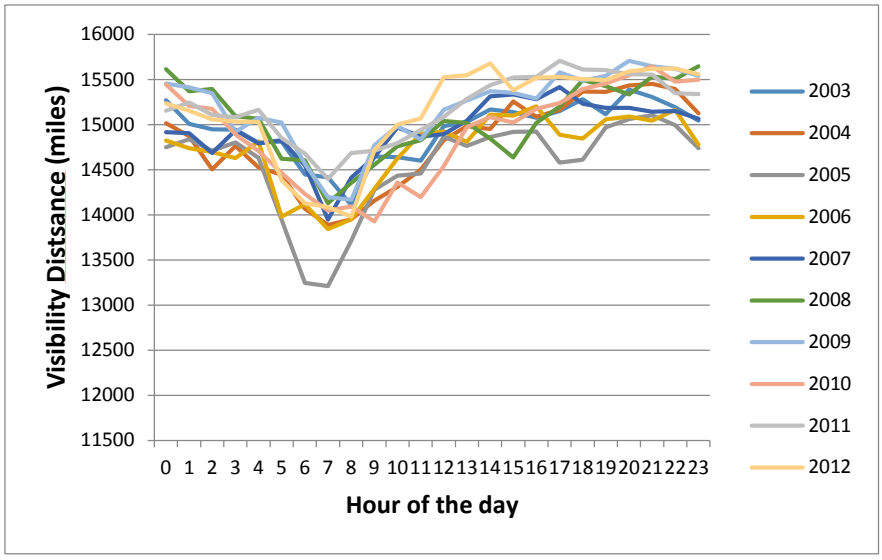

Figure 1: Average visibility across the day at SFO

2. Analyses using flight data solely

Users of the database can also analyze various trends using the flight data alone. In the following example, we looked at the annual trend of flight cancellation across the 10 years, and noticed that there was a dip in October 2012. Further investigation was done to evaluate the causes and impacts of the increased flight cancellation in that month. After narrowing down to specific airports and dates, we were able to cross-check the dates and attribute the cause of the high cancellation rate to Hurricane Sandy. From Figure 2, it could be observed that impacts on flight operations were mainly limited to the Northeast region of the country, and that Southwest Airlines underwent a faster recovery after Hurricane Sandy.

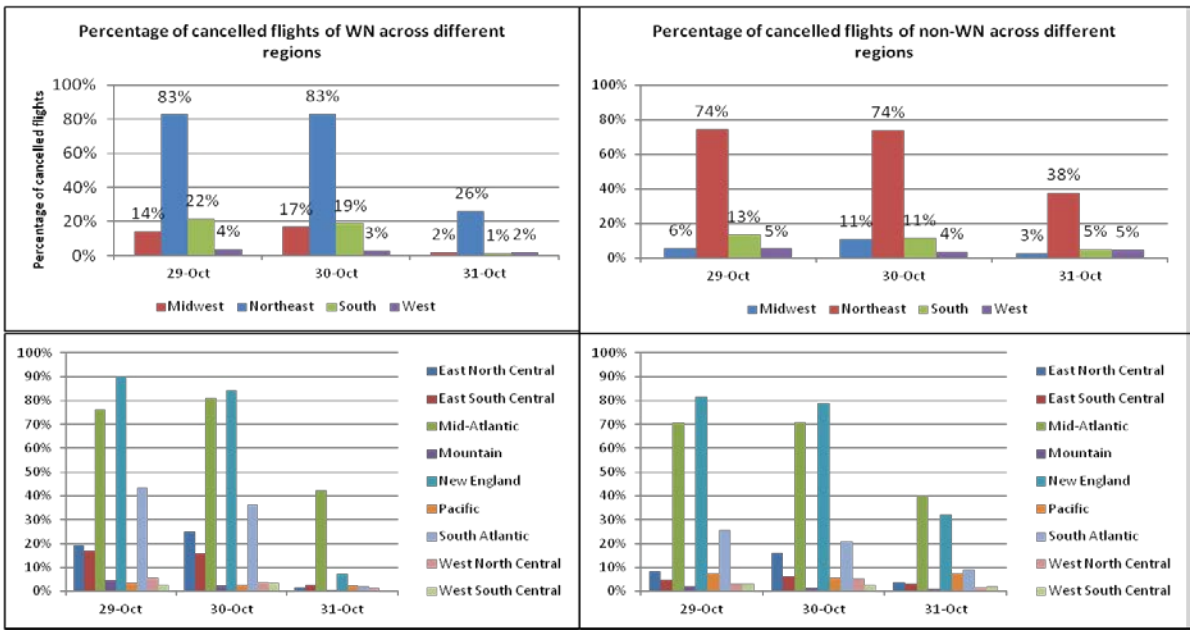

Figure 2: Percentage of cancelled flights across the country during Hurricane Sandy

3. Analyses merging both flight and weather data

Last but not least, by using SQL JOIN statements, users can link up the two tables to examine the correlation of various weather factors and flight performance at any major airports. Figure 3 below shows the delayed percentage at different ranges of visibility distances at the William P. Hobby Airport (HOU) in 2012. Prior analyses and literature research revealed that sea fog in the early morning hours is a common weather phenomenon at Houston.[6] Thus, we investigated further into the effects of sea fog on flight operations. By plotting the average percentage of delayed flights at different ranges of visibility distance, we can clearly see a negative correlation between visibility distance and the percentage of delayed flights. It should also be noted that the proportion of delayed flights increased sharply as visibility dropped below 3 miles.

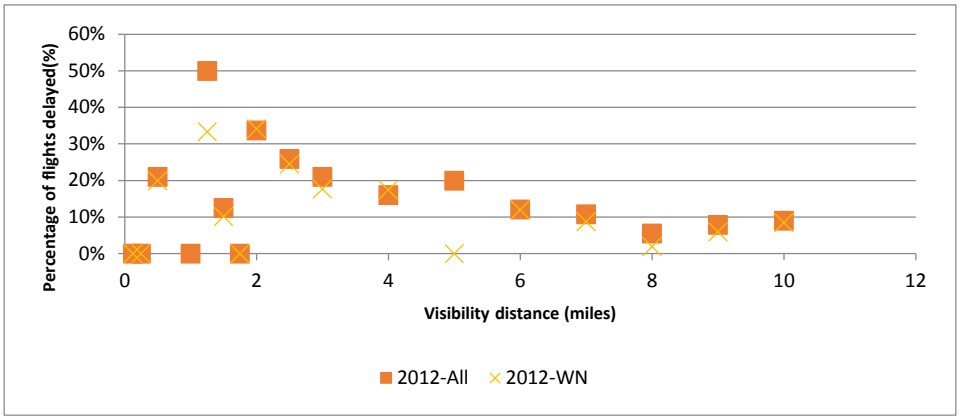

Figure 3: Arrival on-time performance against visibility distance in 2012 at HOU

[6] Atkinson, S. (n.d.). *The effects of wind direction on Houston weather*. Retrieved from http://www.theweatherprediction.com/weatherpapers/078/index.html

IV. Future Work

The ultimate goal of the weather index project is to develop a dynamic and user-friendly tool that forecasts daily on-time performances given certain levels of weather factors. The tool aims to assist airlines in their recovery decision-making and to minimize the extent of flight delay propagation, especially in face of critical weather conditions. To achieve this goal, we identified two key areas, which we hope to further explore in the next step of our research.

i. Identifying baseline levels -
   Baseline levels should be determined in both the weather aspect and the flight on-time performance aspect. In terms of on-time performances, the percentages of delays and cancellations due to non-weather related reasons, including but not limited to, aircraft mechanical problem, air traffic congestion, crew availability, flight scheduling, would form our baseline model. On the other hand, baseline readings for weather conditions would be the levels at which most flights could land with the potential of no delays.
ii. Quantifying system-wide effects -
   By quantifying system-wide effects, we hope to be able to predict how the entire national aviation system would respond to unfavorable weather at airports in different regions. Additionally, the analysis could further be refined to distinguish the impact of severe weather conditions has on a major hub and that on an outstation.

With this tool, we hope that airlines can better optimize block scheduling and crew scheduling based on given weather predictions, so as to reduce the frequency and extent of flight delays, ultimately offering better services to their passengers and eliminating unnecessary costs incurred.

V. Acknowledgement

I would like to express my deepest appreciation and thanks to my faculty advisor, Professor Amy Cohn, for offering me this research opportunity and for her continuous guidance and support over the past few months.

In addition, I would like to extend my sincere gratitude to Mr. Tony Wang for his generous financial support that enabled me to participate in this research project, as well as Mr. Mikhail Zolikoff, Regional Director for the Southwestern States for the Office of Advancement in the College of Engineering, for his guidance.

Special thanks must be extended to Mr. Rodney Capps, Mr. Gene Kim, and Mr. Christopher Konrad in the Industrial and Operations Engineering Department, as well as Mr. Andrew Caird, Director of High Performance Computing in the College of Engineering, and Mr. Brock Palen from the Center for Advanced Computing for providing a tremendous amount of vital computing advice and assistance throughout the project.

Last but not least, I would like to acknowledge Mr. Eric Camacho, Mr. Lonny Hurwitz, and Ms. Ann Nguyen from the Network Planning Team of Southwest Airlines, as well as the following students who offered much support and made valuable contributions in this project: Mark Grum, Sanjeev Muralidharan, Donald Richardson, Luke Simonson, George Tam, Zachary Verschure and all other CohnAPalooza2013 students.

Without the help and support of the aforementioned people, the project would not have moved along as smoothly as it had.