Exploratory data analysis of US airline travel for January 2016

Amine Laghaout
(Dated: October 10, 2016)

Executive summary: This cursory report provides a few insights about domestic flights within the United States for the month of January 2016 [1]. Delay times by airports and by carrier are quantified, and their fluctuations over a 24-hour period are reported. Of particular relevance to the punctuality-conscious traveler, a recommendation of airlines is provided for specific links between the busiest airports.

Keywords: airline travel; punctuality; airport delays; inter-city recommendation

Contents

I. THE SCOPE

This emphasis of this report is descriptive rather than predictive. This is partly due to the time constraints, but mainly to the nature of the data which revolves almost exclusively around the time performance of carriers and airports. One could have conceived of a predictive analysis which, for example, could classify airlines as prone to delays (or not). However, the lack of other non-categorical data (e.g., passenger occupancy, price of tickets, etc.) makes it much easier to limit ourselves to exploratory analysis.

The main quantitative figure of merit under consideration is punctuality (i.e., small delay times). Delays can be incurred by inefficiencies in the operations of either airports or airlines [4] We shall therefore address these two factors separately. In Sec. II, we shall look at the ten busiest airports—in terms of overall flights—and rank them in terms of taxi time. In Sec. III, we perform a similar analysis of arrival delays by airline. Then, in Sec. IV, we propose a table which could be used by travelers as a reference for which airline to choose so as to minimize arrival delays. Finally, in Sec. V we present the fluctuation of arrival and departure delays over a 24-hour period so as to visualize the busiest times of day.

The data was analyzed using Python 3.0 and its `Pandas` library [2]. The source code is attached as `main.py` and `module.py`.

II. AIRPORT TAXI TIMES

This section looks at the taxi times—both taxi-in and taxi-out—for the ten busiest airports over all airlines that serve them. The mean, median, and 1st and 3rd percentiles of these taxi times are visualized in Fig. 1. The main conclusion is that, among the ten busiest airports, Chicago O'Hare fares worst, with an average of 15.2 minutes of taxi time whereas Las Vegas' McCarran International is most efficient with only 11.3 minutes.

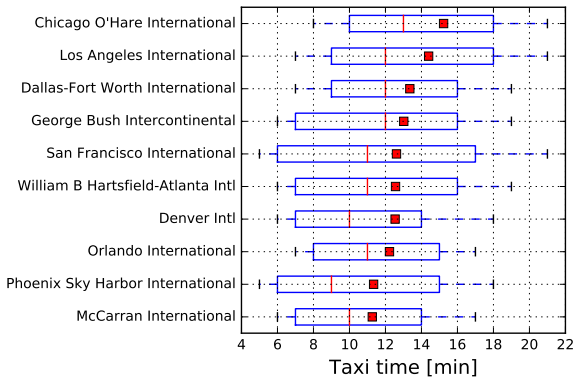

FIG. 1: Box plot of taxi times (both taxi-in and taxi-out) for the ten busiest airports (as per number of flights). The mean is marked with a solid red square and the median by a red line. The whiskers extend from the 15th to 85th percentile.

The reason for choosing *taxi times* rather than overall departure or arrival delays is that taxi times are (arguably) more intrinsically linked to airport operations, whereas other types of delays are much more prone to airline operations (e.g., passenger boarding, luggage handling, etc.).

III. CARRIER PUNCTUALITY

As was just mentioned at the end of the last section, the punctuality of an airline—though dependent on the efficiency of the airports it operates from—is best summarized by the overall delay time of *arrivals*. (Departure delays are less relevant in that they are included in arrival delays.) This is shown in Fig. 2 where it tran-

spires that among the ten-busiest airlines (per number of flights), Spirit Air Lines is the most prone to delays with an average delay of 13.1 minutes after the scheduled arrival times. A surprising finding, however, is that the five major airlines actually arrive earlier than their scheduled times with Alaska Airlines arriving almost 4 minutes early on average. This is corroborated by an earlier finding on airline punctuality published in USA Today on January 7, 2016 [3].

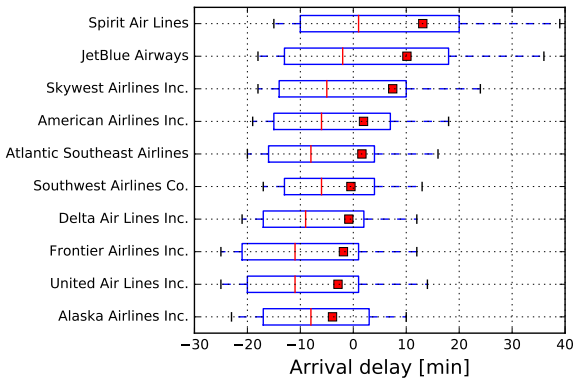

FIG. 2: Box plot of arrival delays for the ten busiest airlines (as per number of flights). The mean is marked with a solid red square and the median by a red line. The whiskers extend from the 15th to 85th percentile.

IV. INTER-CITY RECOMMENDATION

This section is likely to be of most utility to the frequent passenger who wants to select the most punctual airline linking any two of the ten major airports. Table 3 lists at the intersection of departure (columns) and arrival (rows) airports the most punctual airline available in the month under consideration. For example, if one were to travel from Orlando (MCO) to Chicago (ORD), then United Airlines (UA) is the carrier that has the smallest delay.

It is interesting to note that the table is asymmetric. For example, although it is preferable to travel from Orlando to Chicago with United Airlines, the other way around is better served by American Airlines. One plausible reason for this asymmetry is the fact that Chicago is the hub of United Airlines and that its aircrafts have a preferential treatment on the way in. However, this claim needs further evidence to be validated.

V. DELAYS OVER 24 HOURS

This last section looks at the arrival and departure delays over 1-hour time blocks from 6 am to midnight in addition to another time block from midnight to 6 am. This is summarized in Fig. 5. The two conclusions that can be drawn from this are that arrivals are extremely fluid from 6 am to about noon whereas both arrival and departure delays rise constantly over the daytime to peak at around 7 pm.

VI. A NOTE OF CAUTION

As mentioned in the abstract, this report is very *cursory*. The main drawback of all the results presented herein is that they collapse several dimensions and only look at marginal distributions, thereby averaging away potentially relevant information. For example, Fig. 5 is produced over all days of the week and thus does away with the important fluctuations that differentiate weekdays from weekends. Similarly, the recommendations of Table 3 consider the whole month of January and will therefore be of no use to the passenger that needs to travel a particular day when the recommended airline does not operate.

[1] B. of transportation statistics, *On-time performance data*, URL `http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time`.

[2] P. Community, *Python data analysis library*, URL `http://pandas.pydata.org/`.

[3] U. Today, *Report reveals the best airlines and airports for on-time performance*, URL `http://www.usatoday.com/story/travel/flights/todayinthesky/2016/01/07/airline-on-time-performance/78407958/`.

[4] A much more likely scenario is that the "joint inefficiencies" of airports and airlines are at play: The disentanglement of which one of these two players lies at the root cause of any given delay may require much deeper analysis such as Independent Component Analysis (ICA).

	ATL	DEN	DFW	IAH	LAS	LAX	MCO	ORD	PHX	SFO
ATL	n/a	UA	DL	EV	WN	AA	DL	EV	DL	UA
DEN	UA	n/a	UA	F9	UA	UA	UA	OO	UA	OO
DFW	DL	UA	n/a	EV	AA	AA	NK	AA	NK	UA
IAH	UA	UA	EV	n/a	F9	UA	UA	UA	UA	F9
LAS	WN	OO	AA	F9	n/a	UA	WN	AA	WN	OO
LAX	WN	F9	OO	UA	UA	n/a	UA	AA	UA	UA
MCO	WN	UA	AA	F9	F9	F9	n/a	AA	WN	UA
ORD	UA	F9	UA	UA	F9	UA	UA	n/a	UA	UA
PHX	WN	F9	AA	UA	WN	UA	AA	AA	n/a	UA
SFO	UA	WN	UA	F9	UA	UA	UA	F9	UA	n/a

ATL	William B Hartsfield-Atlanta Intl
DEN	Denver Intl
DFW	Dallas-Fort Worth Intl
IAH	George Bush Interc
LAS	McCarran Intl
LAX	Los Angeles Intl
MCO	Orlando Intl
ORD	Chicago O'Hare Intl
PHX	Phoenix Sky Harbor Intl
SFO	San Francisco Intl

FIG. 3: Table of recommendations linking departure (columns) and arrival (rows) airports. Only the ten most traveled airports are listed. Note that the table is asymmetric, presumably due to the fact that some airports serve as hubs for certain companies. The look-up table for the airport codes is provided in Table 4. For obvious reasons, the links from any given airport to itself, are marked as not applicable (n/a).

AA	American Airlines
DL	Delta Air Lines
EV	Atlantic Southeast Airlines
F9	Frontier Airlines
NK	Spirit Air Lines
OO	Skywest Airlines
UA	United Air Lines
WN	Southwest Airlines

FIG. 4: Look-up table for the carrier codes used in Table 3.

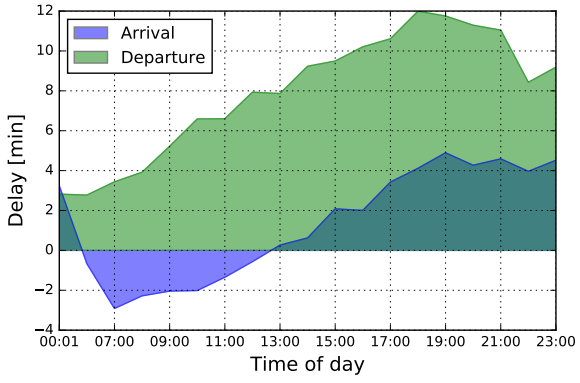

FIG. 5: Mean delay time for both arrivals and delays over all airports under a 24-hour period. Note that all times correspond to 1-hour time blocks except for the first one, from midnight to 6 am, which is collapsed into a single time block.