Critical Review on the FiveThirtyEight article on the Best and Worst Airlines

- *Karthik Abinav Sankararaman*

The article "The Best And Worst Airlines" uses innovative methods to rate the various airline providers and the airports based on delays and cancellation. It distinguishes itself from the statistics reported by the government in the fact that, it also accounts for certain key offsets. The airlines **pad** additional time in their schedules to account for unseen delays. The government statistics uses the deviation from the scheduled time reported by the airlines for computing the average delays. On the other hand, this article uses a different baseline, which it claims to be **fairer**. The article measures average travel time compared to other airlines flying the same route. This article falls into the category of "**Data Journalism**" as defined in Coddington's paper. The main theme of the article is to report about the performance of various airlines supported by data

analysis. Let us delve deeper into their analysis and see the pros and cons of their approach.

The central argument of the article is to **rate airlines** based on their average delay in a chosen period, where per flight delay is normalized by its source and destination airport delays. For example, Chicago airport has a much higher average delay as opposed to Honolulu because of the weather conditions. Hence, a flight departing Chicago is given some additional leeway as opposed to Honolulu. This leeway time is calculated based on average delays for the airports available from previous statistics. Additionally, their statistics accounts for the faster flying time for eastwards flight as opposed to the corresponding westwards flight by having separate flight times for each direction. Hence, the article defines the notion of **target time** for a flight from destination A to destination B. Actual flying times are compared to this baseline in their analysis. Though this is a detailed and careful choice of baseline, in my opinion, they missed out a couple of important

factors. The first factor is not all flights going from destination A to destination B travel the exact same route. Traffic congestion leads to scheduling issues and usually ATC is responsible for allocating a particular air space to a particular flight (American Airlines [1] versus Delta [2]). Hence, intermediate weather conditions play a huge role, which the article does not consider in its analysis. The second important criteria, which the analysis does not consider is that different models of aircrafts take different flying times under the same conditions. For example, Boeing 777 has a top speed of 590mph [3] while Boeing 747 has a top speed of 614mph [4]. Hence, for a particular route these criteria should also be accounted for. Note that, the dataset provides this information for every route.

The data used in the article is the one collected by the **Bureau of Transportation Statistics** [5]. The article uses the data for Summer 2015 collected for the period May 2014 to April 2015. The data, for most parts is official and reliable since this is obtained

directly from the Air Traffic Control (ATC) at various hubs. The analysis is done for about 300 major hubs within the US. ATC keeps track of arrival time, defined as the time when the aircraft reaches the gate and the aerobridge is connected to the door, departure time, defined as the time at which the aircraft pushes off the tarmac for every flight at that hub. Hence, there is no issue of missing data/values. However, this dataset alone does not suffice for all their calculations. They use the **Google maps API** in certain scenarios, as we will see in the next paragraph. Though, this does not cause the issue of missing values, there might be possibility of a different kind of error, which we will see shortly.

The dataset given by the Bureau has a lot of fields among which the article uses the following fields. They collect the origin airport id, destination airport id, departure time, arrival time and cancelled status. In case the flight isn't cancelled the calculations are straightforward. However, for cancelled flights they assume it takes **three extra flights** before the

customer can be accommodated, on average. So they look for the next three flights from that time and account for that. However, they have a cap of 240 minutes since a serious incident (such as the recent Washington Blizzard [9]) will cause a lot of disruption to a series of flights and it is unfair to charge these incidents as delays. One innovative approach their analysis does is to use the Google maps API to check the driving time between two airports. If that time is lesser than the next available flight, they assume that the passenger will drive to their destination. In my opinion, this introduces two kinds of errors. Firstly, not all airline companies pay for ground transportation. For example, United airlines [6] clearly mentions that they will pay for ground transportation only if they are unable to accommodate them in their hotels. Hence, it is inaccurate to assume that passengers use this option frequently. The second biggest problem is with the maps API itself. “Maps API” provides the information of *current traffic as opposed to traffic at a previously chosen time* [7]. In other words, a major cancellation such as the

Washington Blizzard will have much different traffic patterns (and hence different driving times) as opposed to the one when the analysis was done. Hence, not accounting for this makes this analysis inaccurate.

The article shows a nice visualization of their results [8]. We can select various source and destination airports and see the target time as well as the actual average times for various aircrafts. We can also look at a particular airport and all routes and their delays. The map also colors different routes with three colors indicating performance, categorized as fast, medium and slow. The second visualization shows the delays month-wise for different airliners for this timeframe. The visualization was highly interactive and gave a clean summary of the entire article in a user-friendly interactive way.

Overall, the article was well written with an innovative use of various publicly available dataset.

The main story entirely depended on data and the analysis was transparent, clearly written and had a nice interactive visualization for people interested to delve deeper. One particular aspect I would like to see is another analysis, which looks at this data for a period larger than 12 months. For a given place, a 12 month period accounts for seasonal variations. However, not all winters are equal in terms of weather. Some winters are much harsher (e.g. Winter 2014-2015 [10] vs Winter 2015-2016 [11] in Washington DC) than others. Hence, taking a larger window might average out these effects and might give a stronger evidence for performance. Right now this is partially accounted for, in the form of calculating the leeway times. Hence, a complete analysis will round this article up in a nice way.

References:

1) http://aa.fltmaps.com/en

2) http://www.delta.com/content/dam/delta-www/pdfs/route-maps/us-route-map.pdf

(3) https://en.wikipedia.org/wiki/Boeing_777

(4) https://en.wikipedia.org/wiki/Boeing_747

(5) http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

(6) https://www.united.com/CMS/en-US/travel/policy/Pages/FlightDelaysandCancellations.aspx

(7) https://developers.google.com/maps/documentation/javascript/examples/layer-traffic

(8) http://projects.fivethirtyeight.com/flights/

(9) http://www.usatoday.com/story/travel/flights/todayinthesky/2016/01/23/blizzard-2016-air-travel-nightmare-extends-into-saturday/79220674/

(10) https://www.washingtonpost.com/news/capital-weather-gang/wp/2015/04/09/final-snowfall-total-map-for-2014-2015-another-overachieving-winter-in-the-d-c-area/

(11) http://heavy.com/news/2016/01/washington-dc-snow-snowfall-totals-amounts-winter-storm-jonas-blizzard-2016-snowzilla-by-town-city-weather-service-official/