RTI CDS Exploratory Data Analysis (EDA), National Transportation Safety Board Dataset

Javad Roostaei, Ph.D., June 2020

1. Introduction

I used the Aviation Accident Database & Synopses from the National Transportation Safety Board to analyze the data on accident reports from 1948 to 2015. In addition, I reviewed the 144 JSON files that contain the "narrative" and the "probable cause" of the incidents.

2. Methods

I used python and different libraries supported by python to carry out this analysis. All results can be seen in a Jupyter notebook uploaded in my GitHub account and expressed in three parts. Part1: EDA on XML file, Part2: EDA, and NLP on JSON files, Par3: Analysis of combined XML and JSON files.

3. Results and Discussion

3.1. Exploratory Data Analysis for the XML dataset

After parsing the XML dataset, there are 77248 rows and 31 columns. Two columns present longitude and latitude data, and five columns present numerical data on injuries. The remaining contain categorical string data. Figure 1. presents the results from exploratory data analysis. More visualization and descriptive statistics can be found in the notebook. The geographical distribution, fig1(A), indicates that 94.6% of all incidents were reported in the US and that weather conditions for 89% of the incidents were reported as Visual Meteorological Condition (VMC). The purpose of the top five flights during which incidents occurred, as presented in fig1(B), indicates that the majority took place during personal flights, followed by instructional flights. Similarly, fig1(C) indicates that Landings and Takeoffs were the phases during which the most incidents occurred. On average, the most fatal injuries (fig1(D)) took place in July and August, probably because more flights took place during the summer months. On average, the fewest mortalities occurred in March. It should be noted that total injures (fatal, serious) and incidents, presented in fig1(E), show a marked decline from 1982 to 2015.

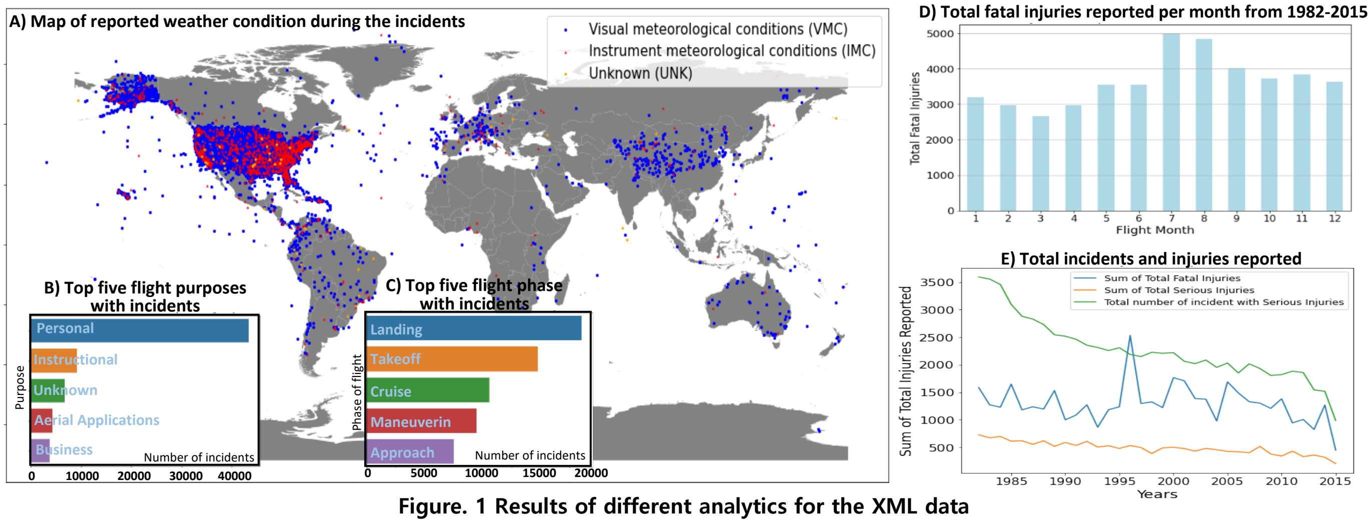

Figure. 1 Results of different analytics for the XML data

3.2. Natural Language Processing for JSON files

Results of NLP for the "narrative" and "probable_cause" columns show that "**pilot, airplane, aircraft, accident, investigation**" and "**pilots, failure, landing, maintain, control**" are the top five words used in each column respectively. I used left join to merge the XML and JSON files based on the EventID and did some additional analysis based on the date. Then I filtered my data for most recent (2010 to 2015) and from 2000 to 2005. The top 20 words used in the narrative column did not change significantly over the two selected periods. Next, I checked for a specific word, i.e., "sleep," which appears more frequently between 2000 and to 2005 than in earlier time periods. More analysis can be found in the Jupyter notebook.

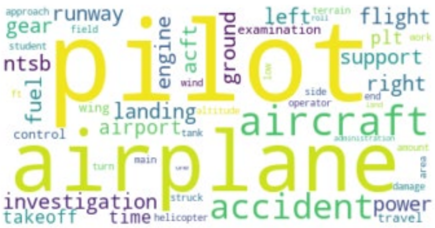

Figure2. Top 50 frequent words in the narrative column

4. Conclusions

In summary, the number of incidents and injuries are declining from 1982 to 2015. Personal flights are responsible for the majority of incidents, suggesting that more regulations are needed to increase the safety for personal flights. Pilot(s) is the most common word used in the "narrative" and "probable_cause." To avoid incidents, landings and takeoffs are the phases that call for more attention and training.

References

1. Aviation Accident Database & Synopses, https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx
2. Data dictionary for processed NTSB summary data, http://www.airsafe.com/analyze/ntsb-data-dictionary.pdf