International Journal of Data Analysis and Information Systems
Volume 9 • Number 1 • June 2017

Transportation Analytics: A Study of Aviation Accidents and Flight Incidents

Madhumita Deepika Duvvuri[1], Sravan Kumar Borra[1], Prithvi Yarlagadda[1] and Sylvain Jaume[1]
[1] Saint Peter's University, 2641 John F. Kennedy Boulevard, Jersey City NJ 07306, USA
E-mails: vduvvuri.student@saintpeters.edu; sborra@saintpeters.edu;
pyarlagadda@saintpeters.edu; sjaume@saintpeters.edu
corresponding author: sjaume@saintpeters.edu

ABSTRACT

We perform an analysis of the global aviation accidents and incidents over the past years. We examine the associated features and their effects leading to the accidents or incidents globally. The number of fatalities and the rate of accidents by an airline were calculated based on 31 different contributing factors in the dataset with dates being an important factor as it allows us to track the patterns and trends. Over the years the National Transport Safety Board (NTSB) identified these accidents and incidents in various countries. We used a database to visualize information related to the factors responsible for the accidents. Visualizing these features allows us to find a probable solution or avoid the accidents and could help in taking measures to decrease the number of accidents.

Keywords: Transportation, database, aviation, flight incidents, time analysis, geographic analysis, analytics, Accident Prevention

INTRODUCTION

Aviation safety is interrelated and complex, just like the aviation system. Safety includes airports and aircraft design, airfields, training of flight crew members' and ground personnel, aircraft maintenance, communication and navigation facilities of en route airlines and terminal area navigation, air traffic control procedures, implementing Federal Aviation Regulations (FARs) and much more. Every aspect involving aviation must include the safety feature. Amongst all the transportation modes, aviation is the only transportation where safety record is rigorously and continuously scrutinized. In fact according to the latest statistics, U.S. aviation transportation is on a safety streak with no fatalities due to accidents reported. This means that no one died in a crash of a U.S. certified airline in 2016 operating anywhere in the world.

Forbes contributor, Dan Reed states in his latest aviation post (December 2016) that "In fact according to the latest statistics, U.S. aviation transportation is on a safety streak with no fatalities due to accidents reported. This means that no one died in a crash of a U.S. certified airline in 2016 operating anywhere in the world. It is the seventh straight year that nobody died in a crash on a United States-certificated scheduled airline operating anywhere in the world".

Many airlines conduct aircraft familiarity classes for passengers who have a sense of fear when they fly, although with experience of the latest technology in air transport this phobia has become less prevalent in the younger generation. Aside from many self-enforcement attempts made by the industry, the Federal government also tries to ensure the safety of the passengers through regulation.

The National Transportation Safety Board (also commonly known as NTSB) is a government agency that investigates all air carrier incidents or accidents and consequently regulates safety norms to the Federal Aviation Administration (FAA). The FAA may or may not choose all these safety norms to be included in its recommendations. One of the longest conflicting ideas in aviation safety is between the NTSB (whose sole purpose is safety) and the FAA, which also considers safety regulation economics into account, otherwise for every change in FAR there

will be a chain of industry reactions. The International Civil Aviation Organization, also known as ICAO, (International Civil Aviation Organization, 2017) which investigates accidents on an international level regulates technical rules which affect aviation safety, although these decisions will be taken considering the economic conditions as well.

The safety regulations in the aerospace industry (external or internal) should make some sense in reality to some degree, i.e. every regulation must be supported by statistical, technical, or economic factors which can be addressed later on their own qualities. Safety regulations are successfully implemented and accepted by the industry if the changes or rules are supported by quantitative data. Aviation safety analysis thus came into subsistence, which mostly means that the aim of safety analysis is to increase efficiency and improve safety. Broadly we can say that the analysis spectrum ranges from the investigative analysis to the predictive analysis. One end of this spectrum will be the investigation of accidents or incidents and explore for causes and at the other end of the spectrum is the effort to know the likely causes (or, combination of causes) of failure. Predictive analytics on aviation data is tested in various studies (Bineid *et al.*, 2003) and investigative analyses (Han *et al.*, 2001).

However, one of the greatest difficulties in aviation system analysis is that the data available is not sufficient enough to make probabilistic conclusions even though the main objective of this study is to eliminate or to avoid the very accidents or incidents from which we retrieve the data. Classical statisticians consider probability as the outcome (i.e. likelihood) of an event based on a large number of repeated trials and it is difficult for them to accept the concept that an event that has never taken place can nevertheless be assigned a 0.95 probability of success. This concept is essential in understanding the difference between the investigative and the predictive methods of safety analysis. While the first one is based on few real accidents, the second one is based on subjective probabilities of a system or subsystem failures. Aviation safety analysts have accepted that there is insufficient data after only one accident occurs and simply wait for the next one to happen, so they can combine both investigative and predictive methods to perform and improve safety analysis.

The goal of this paper is to perform data analysis on the global aviation accidents and incidents in the past few years. This analysis primarily focuses on the trends over time, which requires examining all the associated features and their effects leading to the accidents or incidents globally. A main objective of this document was to determine the number of fatalities from these accidents every year in a particular country so that we may be able to prevent these accidents from happening one day.

The number of fatalities and the rate of accidents by an airline were calculated based on 31 different contributing factors in the dataset with dates being an important factor as it allows us to track the patterns and trends. Over the years the National Transport Safety Board (NTSB) identified these accidents and incidents in various parts of the countries. The NTSB is a U.S. government agency that independently investigates civil transportation accidents and incidents. The NTSB is considered to be the lead government agency to investigate and report a civil transportation accident. They conduct these in all modes of transportation like aviation, surface transportation, marine and assistance to different domestic agencies.

Another primary idea of the exploration was to use the database to visualize information related to the factors responsible for the accidents. The features considered are weather conditions (VMC – Visual Meteorological Conditions, IMC – Instrumental Meteorological Conditions), engine type (reciprocation, hybrid rocket and electric), aircraft category (airplane, helicopter, and glider) and phase of the flight (cruise, landing, takeoff, approach, taxi). Visualizing these features allows us to find a probable solution or avoid the accidents and also facilitate in decreasing the rate of accidents. The aviation accident and incidents data is available on NTSB database (National Transportation Safety Board, 2017). This database is regularly updated when an accident is reported. A preliminary report with all the available details is made available online just a few days of the accident or incident (Bureau of Economic Analysis, 2017).

In our paper we use exploratory data analysis and visualizations to understand the aviation safety using the data from accidents and incidents mentioned in the NTSB database. Exploratory data analysis usually means using investigative methodologies that depend on data characteristics like accidents and incidents. The FAA, NSB, NASA, ICAO, airlines, aircraft manufacturers, etc., maintain different database types, which are mostly incompatible to each other. Another complication is that some databases are manual and others are computerized (i.e. different database management systems are used). To create and understand broad trends with this incompatibility is a problem. If the focus of exploration is narrow (for example, a failure of a mechanical part on a specific aircraft), then it might be possible to extract information from the different databases to find a definitive cause. To overcome this problem we have considered the database from NTSB which records the data related to accidents and incidents in the air transport system.

Accidents are caused mainly due to interactive factors like weather, aircraft engine failure, human (the pilot) error and many other factors. Very rarely we have incidents caused by design induced crew errors. Accidents that involve human performance are highly unusual and it is the duty of the analyst to keep this as his main objective so that safety is improved. It depends on the analyst to uncover the sequence of events that lead to the incident or accident even though data unavailability and incompleteness is one of the major drawbacks. It should be noted that the role of defect report systems should be to spot the mechanical failure before they become accidents and the incident report system by humans must be designed in such a way that humans must confess all and every incident so that the safety analyst can isolate the potential incident trends before they become accidents.

NASA manages a reporting system known as Aviation Safety Reporting System (ASRS) to update its database with all the incidents which happened or likely to happen, due to mechanical faults or human error (Solomon, 2006). The main purpose of this reporting system is to discuss and understand the different aspects of safety analysis of general aviation to public transportation system so that they can make some valuable recommendations to aviation safety.

Our analysis explains the global trends and patterns of aviation accidents in comparison with U.S. and Canada. We will further discuss the various features like weather, make of the aircraft, phase of the airline, engine type and others which form the contributing factors in resulting in accidents and in turn causing fatalities due to the crash. Many analyses have concluded that private flying is considered as a self-imposed hazard and is on par with motorcycle racing, mountaineering, and rock climbing (Stratton, 1974). Of the usual factors some 30,000 general aviation accidents have occurred from 1971 to 1977 and are reported in the NTSB database, out of which 67% are due to human error factors, 14% due to weather conditions, 7% due to the engine failure, 6% due to the airport facilities and communication, 3% due to the airframe and the rest to other miscellaneous causes. Similarly, the Federal Aviation Administration reports incidents in the Accident Incident Data System (AIDS) (from April 1983) and recorded 32,712 incidents, out of which 20,319 (62%) were related to human factors meaning that it is the pilot's fault, rather than the machine failure. Many of the accidents occur due to the false perception of the pilot in taking a decision when an incident happens.

The major contributing factor in aviation safety is the human factor and perception of risk by the pilot at the time of the incident. It is always considered as a safety issue on what the effect would result from the pilot's judgment and attitude. Most of the contributing factors in human errors are lack of skill, inattention, carelessness, lack of self-discipline, and many kinds of mistakes. Thus the pilot's attitude is considered as a crucial factor in aviation safety and also equally important is learning these piloting skills and maintaining efficiency. To avoid aviation accidents it is highly important that the pilot adhere to the safety regulations imposed by various agencies like NTSB, FAA and NASA. There are many variables we should consider when we analyze the aviation data and these are described in the data section.

DATA

Global aviation safety data usually involves volumes of data and reliability of the variables noted. This

implies that analysis of aviation safety data entails to set limitations and boundaries to the dataset. So, interpretation of any kind of analyses should be within these boundaries and limitations applied on the data. Similar analyses comparisons where datasets are bounded differently sometimes emphasize interesting issues. For instance, the statistics obtained from Boeing cover accidents which involve Western Built aircraft (over 60,000 lbs). These types of analyses involve most commercial jets but cannot provide a proper reasoning into the safety performance of the aviation systems. Perhaps the analysis of aviation safety data and the way these analyses are presented is generally influenced by the issues of the day or other factors. Therefore, it is not possible to compare different statistical presentations created in the present day with those created few decades ago. This again can reveal the fact on how the public, regulators and industry understand and view performance measured in safety metrics. All of this discussion means that when statistics from different datasets are looked into, it is important to note the boundaries and limitations applied to these datasets and consider what the statisticians have achieved through their analyses.

The data provided by NTSB Aviation Accident database forum (National Transportation Safety Board, 2017) is mostly categorical and the number of accidents, fatal accidents and fatalities are numerical. Categorical data can be defined as a form of discrete data that describes some characteristic or attribute of that particular data. In most of the aviation data, the feature variables describe many attributes which range from what was the engine type of aircraft which crashed, where the accidents happened i.e. latitude and longitude, whether it was a fatal accident or not, if fatal the count of fatalities, what type of weather conditions were involved. These variables from the raw data were reconstructed into different categorical groups to get a better sense of the featured variables when we perform exploratory statistical analysis on this data. With the use of statistical tools like R and Python, we have created a filtered view of the data which had the data required in a particular format and can be used in the process of analyzing the data in an easy manner.

Data collection was classified into two parts for further analysis. The first part will be the data related to number of fatal accidents for U.S. and Canada and the second part will contain the data related to all other nations. The dataset was imported to the database software package Microsoft SQL Server and using this software we will be able to create and edit the large databases. Complex queries can be performed on these databases to retrieve appropriate data for exploratory data analysis. The dataset from NTSB Aviation Accident database (National Transportation Safety Board, 2017) was huge which covered accidents from 1962 to as recent as 2017, and using SQL we collected data from various databases to get the data that was required for this study. The final dataset has around 79700 data entries, including all the accidents and incidents occurring from 2000 to 2017 globally.

As discussed above every dataset is bound to consider some limitations and boundaries are set. Adhering to these regulations the data is populated in the databases. Different data boundaries and limitations are listed below for the reference of exploring the data. These are referenced from the Aviation Safety Performance Reports and Statistics (SKYbrary, 2016).

Type of flight

Aviation data used for statistical analysis purposes usually include only those events that happen when an aircraft undertakes flights for the purpose of commercial aviation i.e. to carry travellers, cargo or mail.

Injury Sustained

Fatality - death occurring consequently upon an accident is classified by ICAO (as such if it occurs within 30 days of the accident) and this definition is usually accepted by others agencies.

Serious Injury - injury occurring consequently upon an accident or a serious incident is normally classified as serious if it results in hospitalization for more than 48 hours that commences within 7 days of the event.

Nature of Event

The following classifications are typically employed:

– Accident" - as defined in ICAO Annex 13. (National Transportation Safety Board, 2017)

An occurrence associated with the operation of an aircraft that takes place between the times any person boards the aircraft with the intention of flight until such time as all such persons have disembarked, in which:

a) a person is fatally or seriously injured as a result of:

— being in the aircraft, or — direct contact with any part of the aircraft, including parts which have become detached from the aircraft, or — direct exposure to jet blast, except when the injuries are from natural causes, self-inflicted or inflicted by other persons, or when the injuries are to stowaways hiding outside the areas normally available to the passengers and crew; or

b) the aircraft sustains damage or structural failure which:

— adversely affects the structural strength, performance or flight characteristics of the aircraft, and — would normally require major repair or replacement of the affected component, except for engine failure or damage, when the damage is limited to the engine, its cowlings or accessories; or for damage limited to propellers, wing tips, antennas, tires, brakes, fairings, small dents or puncture holes in the aircraft skin; or

c) the aircraft is missing or is completely inaccessible.

- Major Accident – An accident in which any of the following conditions is met: The aircraft was destroyed; or there were multiple fatalities; or there was one fatality and the aircraft was substantially damaged.

- Fatal Accident - An accident that results in at least one fatal injury, where death was not due to natural causes or self inflicted injuries, or injuries inflicted by other passengers, and was not due to a malicious act such as terrorism.

- Hull Loss – An aircraft is totally destroyed or assessed to have been damaged beyond economic repair. Assessment as a hull loss is always affected by the age (measured in any or all of years-since-new, cycles flown or landings made) of the damaged aircraft and sometimes by the concern of the operator to avoid the 'public declaration of a hull loss.

- Total Loss/Constructive Total Loss - Statistical data which originates in the insurance market is traditionally a very reliable source of data. Insurers use the terms "Total Loss" and "Constructive Total Loss" which is not quite the same as Hull Loss.

Geographical Boundary

– ICAO Regions are the most often used regional definition. The assignment of region to an accident may be the based on the location of the occurrence or on the state of the Operator as defined by their AOC.

– There is a particular difficulty with the 'definition' of Europe, which may include, amongst other options ECAC, EU, EASA Member States or JAA. Political and regulatory evolution in Europe means that these definitions have themselves appeared, disappeared or varied over time.

Data Limitations

– Extent of Aircraft Damage: Substantial Damage, usually taken as damage or structural failure which adversely affects the structural strength, performance ot flight characteristics of an aircraft and which would normally require major repair or replacement of the affected component(s).

– Aircraft Weight: For fixed wing aircraft, the main distinction employed is between jets and turboprops. In both cases. 5,700 kg / 12,500 lb is commonly used as a lower limit for inclusion in statistical databases. However, ICAO have now (2014) begun to include statistics for aircraft below 5700kg.

– Origin of Aircraft Design: Western/Eastern-built Aircraft - some statistics make this distinction or exclude the latter altogether because data for many operations of Eastern-built aircraft have historically been unavailable or unreliable.

METHOD

We generally use data mining techniques to search for valuable, nontrivial and new information from

large quantity of data (Jiawei *et al.*, 2011). It is usually a cooperative work by computers and humans. We tend to achieve better end result when we can balance between the knowledge of analysts in describing and understanding the problem statement and objectives with which the search capabilities of computers work. Practically, the major purpose of data mining is that we need to be able to predict and perform description analysis from the dataset. Predictive analysis uses some variables in the dataset to predict future or unknown values of another variable of interest. Descriptive analytics mainly emphasizes on finding trends and patterns while describing the data in the datasets that can be interpreted by the analysts. Thus, we can classify data mining methodologies into one of two groups 1) predictive data mining, where a model of the system is produced describing the chosen dataset, or 2) descriptive data mining, where analysis produces nontrivial and new information based on the chosen dataset. As discussed previously data mining is a spectrum of analysis which has predictive data mining at one end where the aim is to get a working model from an executable code which is later used for classification, predictions, estimate values and other methods. On the other end of the spectrum is descriptive data mining where the purpose is to gain a meaningful analysis system that reveals patterns, trends and relationships between different variables in large datasets like our aviation accident data.

Feature selection is a term used in data mining to illustrate the tools and techniques available for reducing the input variables to a convenient size which helps in processing and analysis. A number of authors have implemented feature selection techniques using a variety datasets by (Liu et al., 2007; Grimaldi et al., 2003). It is critical to use feature selection to build a good model for many reasons. One of the reasons is that it implies some degree of cardinality reduction i.e. it requires a cut-off on the number of attributes that should be considered in building a model. Also, the choice of the attributes selected by the analyst or the tool must pick attributes such that they contribute some meaning to the analysis. For effective analysis it is very important to apply feature selection to our dataset as datasets usually have more in-depth information than what is needed to build a good model. The unwanted attributes or features need to be removed from the dataset before building a model as this might reduce the quality of the patterns or trends because of the noise or redundant features. The noise present in some of the attributes makes it difficult to uncover qualitative patterns. To discover meaningful patterns we can use data mining algorithms that require larger training aircraft accident dataset on a huge scale. Some authors (Mitra et al., 2002) used unsupervised feature selection using feature similarity for pattern analysis. In our analysis we have used different feature selection methods for discovering attributes evaluator. The various feature selection methods are: correlation feature selection subset evaluator, consistency subset evaluator, gain ratio feature evaluator, information gain attribute evaluator, OneR feature evaluator, principal components attribute transformer, ReliefF attribute evaluator and symmetrical uncertainty attribute. We have considered correlation feature subset selection methodology in our analysis.

We have applied many data mining techniques on the aviation accident reports in our paper. Out of the many we discovered that feature selection between different attributes helps us to find the rules and relations about the accidents that resulted in fatality. We used Python and R in preprocessing the data to select the variables that can be used in the exploratory data analysis. This paper focuses on different feature selection methodologies to understand and clean the dataset. Descriptive statistics help us to understand and know our data, describe the data to the users, and explore our dataset for meaningful patterns or trends. Exploring the aviation accident database gives us a chance to understand all the basic assumptions, mark the odd values and avoid any oversights visually.

When we come to aviation accident information it is always better to work on the data previously recorded i.e. real time information from past accidents. This allows us to know the patterns like when the accidents happen most, cause of the accident – weather or engine failure, and many more. In aviation industry an accident or incident is something we cannot predict accurately as every incident or accident has a typical cause and the

circumstances in which it happens is different from the others. The main part in avoiding an accident is dependent on the decisions taken by humans (here the pilots) and therefore it is difficult to predict the behavior of a human at the time of risk. Hence, using descriptive analysis and exploratory data analysis on large aviation accident reports brings out the trends and patterns based on all the available features noted during the reports.

Our dataset had many features which included the event type, event id, place where the event happened, type of aircraft engine involved, if the accident was fatal or non-fatal, purpose of the flight, phase of flight, total fatal injuries, total serious injuries and all other information on the accident or incident which happened. For a better understanding we have taken a subset of data with some interesting features. For initial analysis we have considered each variable's possible responses with the total number of responses. Observing the total number of responses for each variable we discarded few potential subsets, as the amount of data lost is huge in some cases. For example, if we considered aircraft category we only have around 22,500 responses out of 79,300 total accidents and incidents. Using weather condition, the number of responses was considerably high – 77,200 responses which allows a complete representation of our dataset. To understand the patterns associated with fatal aviation accidents correlation analysis was performed using feature selection methods. Some of the features are explained below with reference to their correlation with number of fatal accidents.

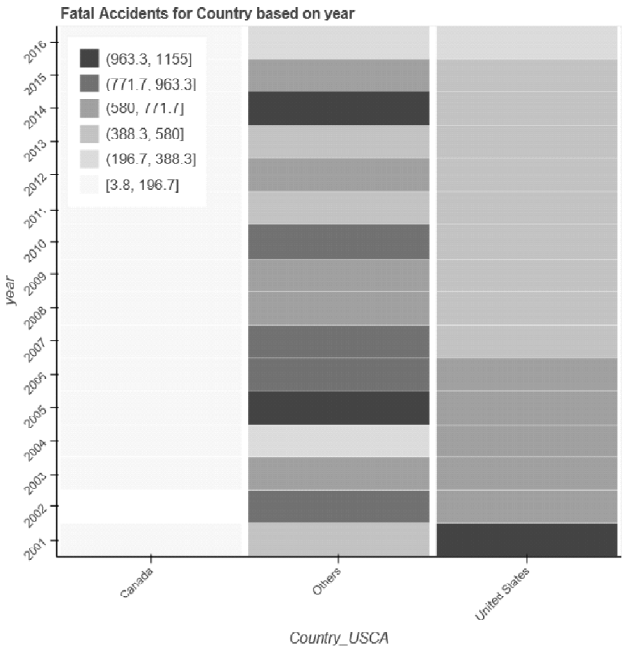

Figure 1: Fatal Aircraft Accidents by Country

The rate of fatal accidents with respect to United States, Canada and the rest of the world from 2001 to 2016 are shown in the heatmap (Figure 1). Observing the graph we can say that the number of fatal accidents in United States has gradually decreased from 2001 and in 2016 the number of accidents resulting in fatalities has decreased to the minimum. Looking at Canada, accidents resulting in fatalities seem to be very low when compared to the rest of the world. Our main concern would be the accidents happening in other parts of the globe that shows alarmingly high numbers with fatal accidents.

Flight Phase is a set of terms used by the IATA and ICAO regulations to classify the operational phase during which an aviation accident or incident happened. It allows a safety trend analysis on occurrences by this category. This taxonomy is classified into two different groups: primary flight phase and secondary flight phase. The primary group has standing, taxi, takeoff, cruise, maneuvering, approach and landing and secondary group has emergency descent, unknown and others. Figure 2 shows a heatmap with number of fatal accidents based on phase of flight we can see that more than one column is clearly darker than the rest of the types, with 'cruise' being highly significant. Cruise is defined as the operational phase of flight after ascent and before descent of an aircraft i.e. this is the phase in which the aircraft is in for the majority of the flight. The remaining darker columns are for approach, takeoff and maneuvering which are the primary flight phases.

Of the many things that are considered as the safety indicators of flight, weather is the most influential and uncertain factor. Weather condition means the characteristics and behavior of the earth's atmosphere at the time of the aviation accident or incident. There are many operational safety issues which can be affected by the weather and some of the noted one's are turbulence, icing, reduced visibility, wind velocity, and many more. A common

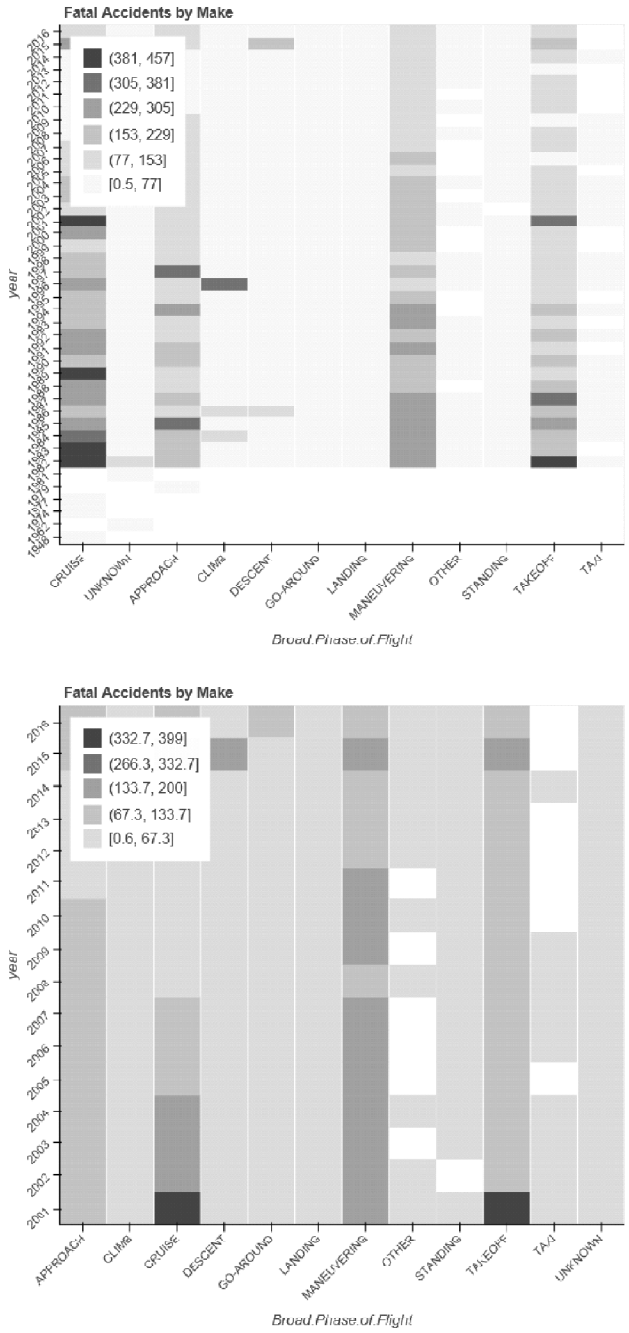

Figure 2: Fatal Aircraft Accidents by Phase-of-Flight

requirement with those associated with the safety of the aircraft should have a clear understanding of the aviation meteorology appropriate to their role.

The terminology used to define weather conditions that caused the aviation accident or incident are VMC (Visual Meteorological Conditions), IMC (Instrument Meteorological Conditions) and UNK (Unknown). VMC is an aviation flight category where the visual flight rules (VFR) is permitted meaning that the pilots have enough visibility to fly the plane while maintaining visual separation from the land and other planes in the vicinity. IMC is another category that details the weather conditions that are required for the pilots to fly primarily with reference to the instruments under instrument flight rules (IFR) rather than by outside visual references under the visual flight rules (VFR). This means that flying a plane is bad weather conditions. Other authors (Nazeri *et al.*, 2002; Shyur *et al.*, 2008; Solomon *et al.*, 2006) performed analysis on aviation data to understand impacts of severe weather on airspace system performance.

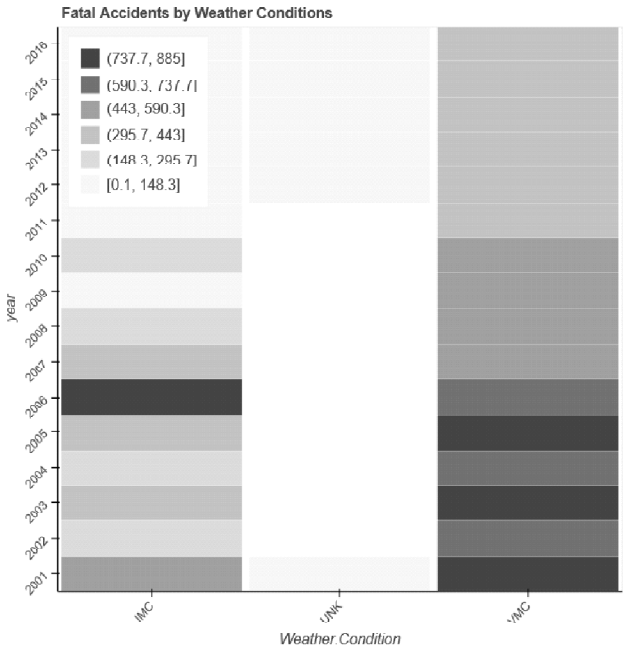

Figure 3: Fatal Aircraft Accidents by Weather

Observing fatal aircraft accident with respect to weather conditions (Figure 3) we can say that accidents where a fatality occurred are when the weather was in "VMC" category. This means that over the years it has been consistent that the accidents that resulted in fatality occurred when the pilot had clear visible conditions and control over the plane rather than when the pilot is dependent on the instrument regulations. From this we can sense a pattern that accidents happening due to VMC weather category are highly consistent over the years. Human factor plays a very vital role when accidents or incidents happen where decisions need to make under critical conditions for the safety of the flight.

Behavioral analysis must be performed on how humans tend to react to a risk situation and what measures we need to take to avoid these conditions (Cruz *et al.*, 2016; Luxhoj *et al.*, 2003).

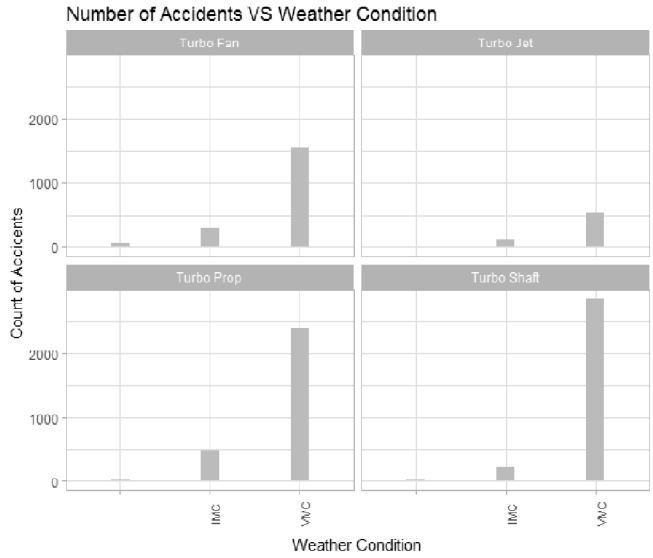

Figure 4: Fatal Aircraft Accidents vs Weather Conditions in specific to engine types

Looking at the heatmap for weather conditions, for IMC the number of accidents with at least one fatality gradually reduced over the years with 2006 noting the highest rate and 2016 as the lowest rate. IMC is when the instrument has the control over the plane. Many analyses related to aviation accidents depicted that the VFR (visual flight rules) flight into IMC accidents were most likely to involve pilots whose experience level is low. There were significant relationships between the accident type and whether the pilot had taken an instrument rating or not. From the analysis we can understand the trend followed in VFR-IMC accidents with pilots who had less training experience (certification) and were least likely to have instrument ratings. From Figure 4 if there was engine failure due to weather conditions and an accident with fatality happened the most likely chances of the crash were due to VMC factor with the pilot having perfect visibility and control over the flight (SKYbrary, 2016).

The year-by-year statistics for aviation accidents is shown in the Figure 5 and it indicates that the number of accidents has reduced drastically over the past few years. The overall number of accidents and the accident rate has reduced which depicts a positive trend for air transportation safety. Globally the rate of aviation accidents has reduced a lot from 1950 with the seventies being the deadliest year for aviation industry when most of the fatal accidents have occurred (Ausrotas *et al.*, 1984).

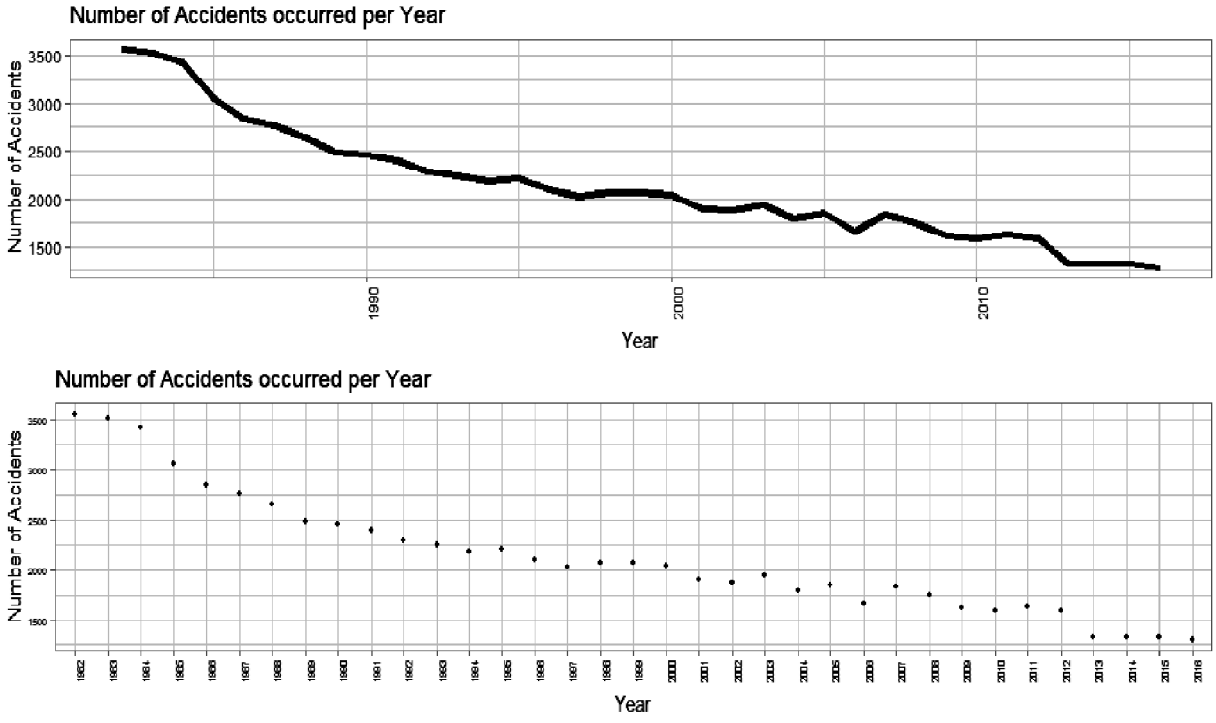

Figure 5: Total Number of Accidents per Year

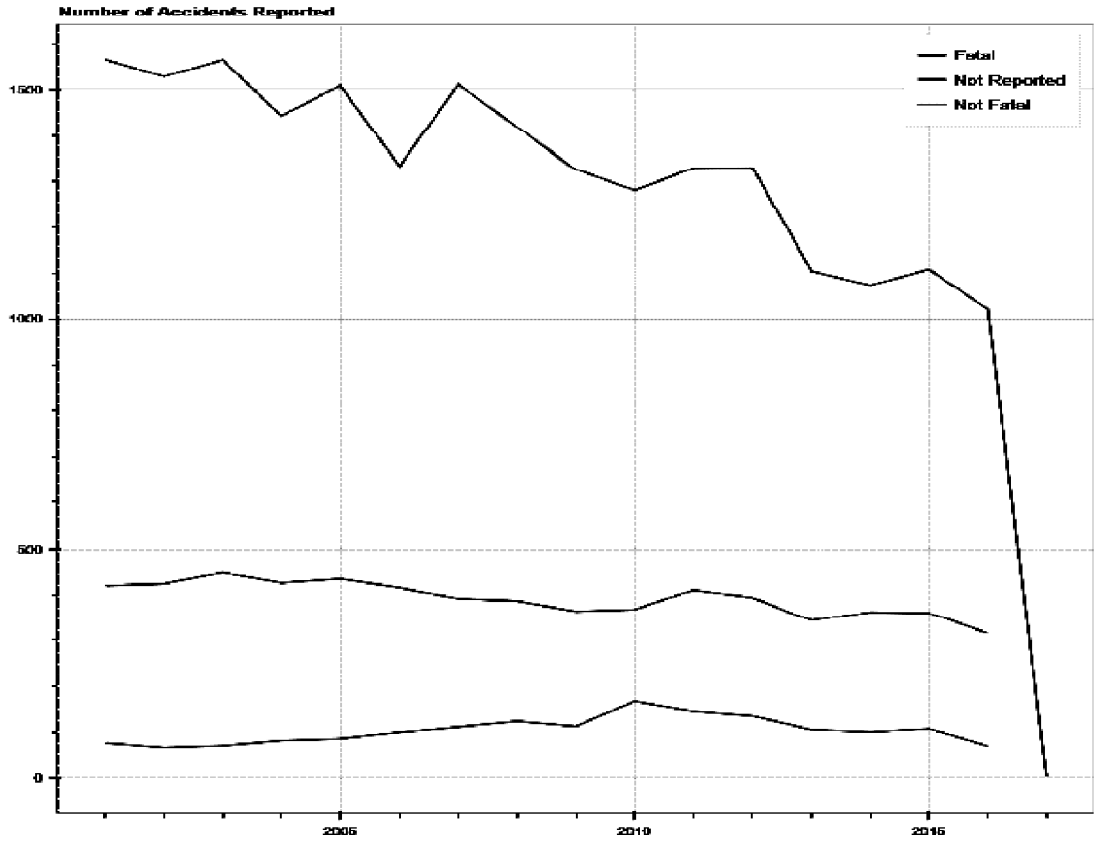

Figure 6: Reported and Non-Reported Fatal Accidents over time

Occurrence of accidents is very rare and consequently the number of accidents occurring over years may vary. So if we focus on one particular year's data then the analysis will be mislead. Considering a 15-year moving average will be useful i.e. the accident rate used is the average of the yearly accident rates over the 15-year gap. The chart (Figure 6) shows that the fatal accidents or accidents in general decrease over the time. Interestingly, the number of accidents is decreasing but the accidents with at least one fatality show a gradual decrease over time. Therefore, evolution of yearly fatal accident rate shows a steady decrease over time and also the total number of accidents has decreased rapidly over time.

RESULTS

Our paper mainly concentrates on the feature selection based on exploratory data analysis and descriptive analysis of aviation accidents and incidents that occurred over a period of 70 years. Observing the various features in the dataset in relation with the number of accidents that result in at least one fatality we have reached to a conclusion that weather conditions are most important when such events take place. Whether the weather conditions resulted in instrument failure or made it difficult for the pilot to take decision is another part of analysis. According to the correlation statistics in exploring the data we have found that United States aviation has been the safest commercial transportation for seven consecutive years than compared to the rest of the world where more number of accidents resulted in fatalities.

Another vital pattern seen is that most fatal accidents have taken place when the aircraft was in 'cruise' flight phase with approach, takeoff and maneuvering phases following the same pattern. Most of the aircrafts with fatal accidents are the one's which are completely damaged. Purpose of the flight is a feature to be considered as it shows the various reasons why the air transportation was chosen as a mode to travel. Most commercial flights have had fatal accidents with purpose of flight being personal

and according to the statistics many safety regulations have been implemented over the years to reduce this rate.

Weather conditions play an important role when it comes to air travel safety, be it directly or indirectly. VMC and IMC are the conditions stated when an aviation accident is reported in NTSB and ICAO. Alarmingly we observe that most fatal accidents occurred when the pilot had proper visible range of the terrain and complete control over the flight. This trend followed here is an assuring fact that either the pilots are less experienced or their behavior is not in accordance to the safety regulations imposed by the governing agencies.

From the year-by-year charts we have understood the trends over time, with the number of accidents decreasing rapidly over the last 70 years but there a steady and slow decrease in the number of fatal accident. Understanding this we have moved to the final step of our exploration with the regions with the most number of accidents and whether fatal accidents are more in those particular regions or they just have incidents reported. Finally we considered the aviation accidents data for 2016 on a global scale to indicate accidents with at least one fatality in red dots and accidents that do not result in any fatality in green dots.

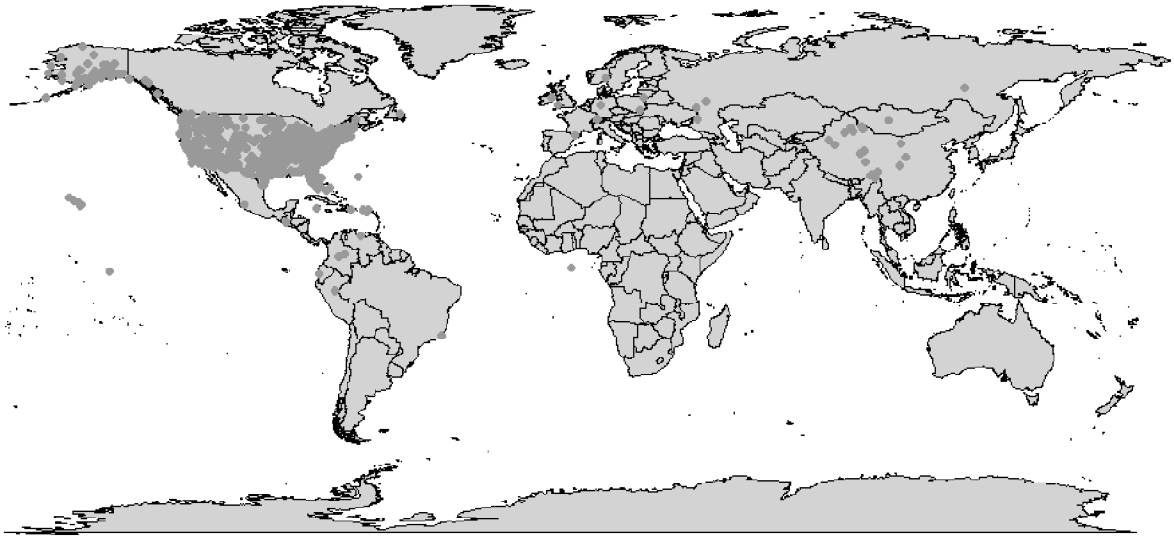

Figure 7: Map of the non-fatal accidents for year 2016 (green dots)

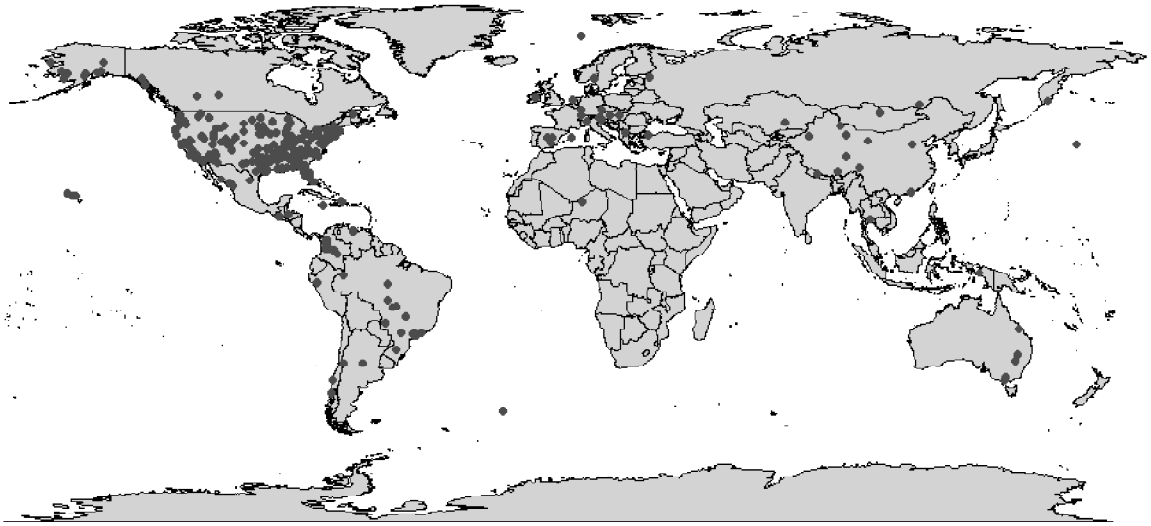

Figure 8: Map of the fatal accidents for year 2016 (red dots)

The geographic map in Figure 7 has green dots for each accident and each green dot represents the accidents that do not result in a fatality. The map in Figure 8 has red dots for every accident and each red dot represents the accidents that resulted in at least one fatality.

Both the graphs pinpoint to all the locations were accidents happened in the year 2016 and count for the total number of accidents is around 1,200. While we can see that most of the accidents or incidents occurred in United States, we do see more of green dots than the red dots. This means that accidents that result in fatalities are less in United States when compared to the rest of the world. Observing the incidents outside the United States, the red dots are slightly more when compared to the green dots which implies that there more fatal accidents which happen outside United States.

CONCLUSION

In conclusion, we can say that in this analysis we explored the large-scale aviation accident and incident datasets using the feature selection methods. We have understood the performance of different feature selection techniques and realized that with any feature selection method we can discover the attributes that tell us the rules and relations about the accidents that result in a fatality. The main contribution of this paper is to perform descriptive analysis and exploratory data analysis on aviation data to identify meaningful patterns, trends and mark the odd values to avoid any oversights visually.

It is always believed that when we discuss any calamity or disastrous event, we find the causes and the reasons why those events happened in the first place. Hence, when we come to the aviation accidents the better approach was to work on the real time data previously recorded i.e. past accidents. This allows us to know the patterns like when the accidents happen most and cause of the accident. Avoiding any accident is majorly dependent on the decisions taken by humans (here the pilots) and therefore it is difficult to predict the behavior of a human at the time of risk. Hence, using descriptive analysis and exploratory data analysis on large aviation accident reports brings out the trends and patterns based on all the available features noted during the reports.

The purpose for our data analysis was to study and investigate the different trends associated with aviation accidents and incidents. The most interesting fact is that the accidents occurred when the weather is clear and visibility range is good enough for pilots to navigate the plane with their expertise and that there are many more incidents happening every year. A positive point about the aviation analysis is that the number of non-fatal accidents outnumbers the number of fatal accidents happening, and on the whole the rate of accidents does appear to be decreasing. One major factor that was not included in this dataset is the cause of the accident. In further analysis, we can combine the results from our analysis with datasets relating the cause of the accident. This can yield more impactful results as to why we still having so many aviation accidents or incidents happening throughout the globe.

REFERENCES

[1] Ausrotas, R. A., and Hansman, R. J. (1984). Aviation Safety Analysis. Cambridge, Mass: Massachusetts Institute of Technology, Department of Aeronautics & Astronautics, Flight Transportation Laboratory.

[2] Cruz, G.S.P., and Sigua, R. G. (2016) Philippine Air Transport Safety: Analysis of Incidents over the Last Two Decades, Aug 8, 2016.

[3] Bineid, M., and Fielding, J. P. (2003). Development of a civil aircraft dispatch reliability prediction methodology. Aircraft Engineering and Aerospace Technology, 75(6), 588-594.

[4] Bureau of Economic Analysis, Industry Data, US Department of Commerce (2017) *https://www.bea.gov/iTable/iTable.cfm?ReqID=51&step=1#reqid=51*

[5] Grimaldi, G., Cunningham, P. and Kokaram, A. An evaluation of alternative feature selection strategies and ensemble techniques for classifying music, Workshop on Multimedia Discovery and Mining, Ireland, 2003.

[6] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2001, Morgan Kaufmann Publishers Inc., San Francisco, California, US.

[7] International Civil Aviation Organization, Economic Development, Aviation Data (2017)

http://www.icao.int/sustainability/Pages/eap-statistics-programme.aspx

[8] Jiawei, H. and Kamber, M. Data Mining: Concepts and Techniques, 2011, Morgan Kaufmann Publishers Inc., San Francisco, California, US.

[9] Liu, H. and Motoda, H. Computational Methods of Feature Selection, 2007, Chapman and Hall/CRC Press, Boca Raton, Florida, US.

[10] Luxhoj, J. T. (2003). Probabilistic causal analysis for system safety risk assessments in commercial air transport.

[11] Mitra, P., Murthy, C.A. and Pal, S.K. Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24, (3), pp 301-312.

[12] National Transportation Safety Board, Aviation Accident Database & Synopses (2017) *https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx*

[13] Nazeri, Z. and Zhang, J. Mining aviation data to understand impacts of severe weather on airspace system performance, Procedings of the International Conference on Information Technology, Las Vegas, Nevada, US, 2002.

[14] Shyur, H.J. A quantitative model for aviation safety risk assessment, Computers and Industrial Engineering, 2008, 54, (1), pp 34-44.

[15] SKYbrary (2016), Aviation Safety Performance Reports and Statistics, Eurocontrol *http://www.skybrary.aero/index.php/Aviation_Safety_Performance_Reports_and_Statistics*

[16] Solomon, S., Nguyen, H., Liebowitz, J. and Agresti, W. Using data mining to improve traffic safety programs, Industrial Management and Data Systems, 2006, 106, (5), pp 621 643.