AN ASSESSMENT OF FLIGHT VARIABLES AFFECTING CIVIL AVIATION ACCIDENTS AND INCIDENTS

Mohammad Kashef

The George Washington University

ABSTRACT

Flight safety has been an important topic for both academia and the industry. Aviation experts and authorities, as well as commercial airline administrators, constantly seek to improve flight safety. Researchers, on the other hand, have tried to model avionic fatalities and suggest improvements or upgrades in flight systems to reduce risk. One approach has been to use data from past accidents and incidents to capture and model the relationship between the different factors involved in each event. However, some important factors are not included in the databases maintained by entities such as the National Transportation Safety Board. This study divides the factors involved into dependent variables (DVs) and independent variables (IVs). IVs include flight factors—for instance, weather and pilot-related data. DVs report the magnitude of the incident/accident, such as the number of casualties. This research will improve existing databases—first, by adding variables, and second, by using multivariate statistical analysis to assess the effect each group of IVs has on correlations between flight factors and accident/incident-magnitude factors. Findings demonstrate that pilot-related factors exert the most influence on the correlation between the two categories. Our findings on the significance of factors or groups of factors will assist researchers, policy makers, flight managers, and flight-crew schedulers in their efforts to increase flight safety.

Keywords: flight accident/incident, causal analysis, multivariate analysis

Dr Mohammad Kashef is currently an Adjunct Professor at the George Washington University teaching systems engineering and engineering management courses for graduate and undergrad students. He has worked as Sr. Systems Engineer for National Oceanic and Atmospheric Administration (NOAA) and has been a member of Joint Polar Satellite System (JPSS) Integrated Product Team (IPT). Email: mkashef@gwu.edu, Phone: +1 703-862-1938

1. INTRODUCTION

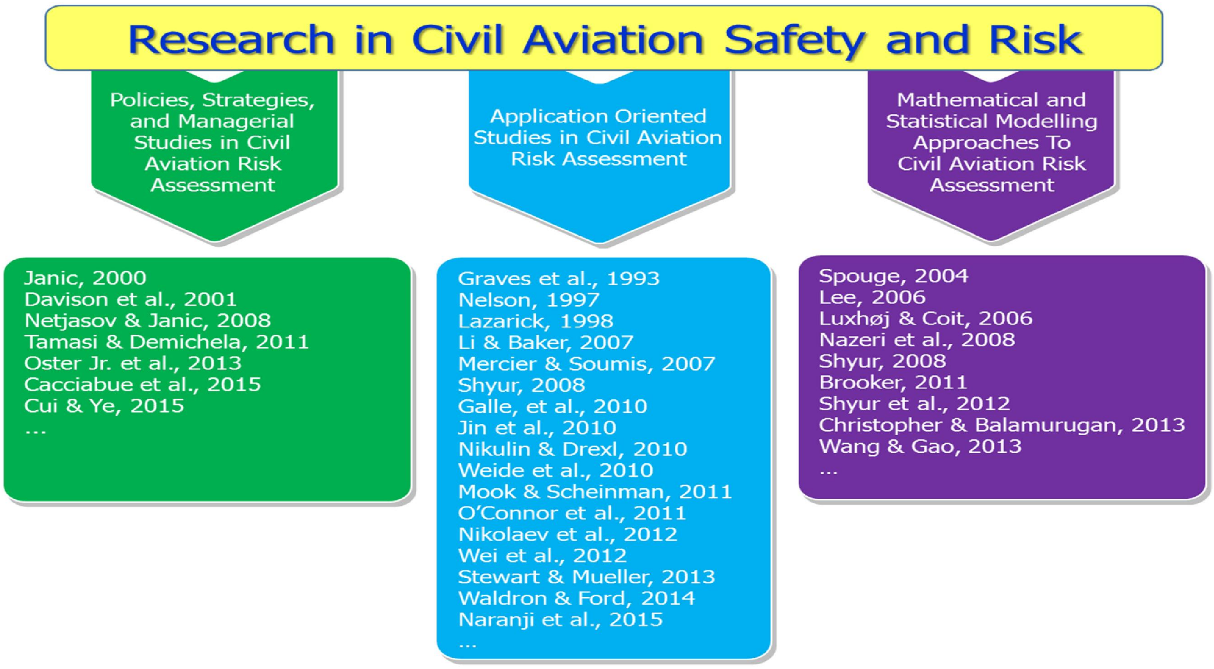
Air transportation is one of the fastest growing transportation modes, with an expected growth rate of 5% to 6% over the next two decades (Netjasov & Janic, 2008). The combination of the complexity of air-transportation systems and their respective interconnectivity with other systems, such as air traffic control and navigation, makes their management highly challenging. Air-transport management, which aims to improve flight safety and reduce the associated costs, covers a broad range of disciplines, from risk management to methods for flight-crew scheduling.

Despite major technological developments in the field, fatal accidents—often with high numbers of casualties—occur with alarming frequency. Recent crashes, and in particular Malaysian Airlines 370 (which disappeared on March 8, 2014, with 239 people on board), Air Asia 8501 (which crashed into the Java Sea during bad weather, killing all 155 passengers), and Germanwings 9525 (which was deliberately crashed by the co-pilot, killing all 150 people on board), have highlighted the critical importance of flight safety.

Aviation events are classified as either 'incidents' or 'accidents' (Nazeri, Barbara, Jong, Donohue, & Sherry, 2008). In an aircraft incident, there are no fatalities, human injuries, and/or substantial aircraft damage; nevertheless, flight safety is compromised. An accident is one in which fatality, human injury, and/or substantial aircraft damage occurs.

Because of its severe consequences, aviation safety has become an important research topic in the past decade (Orasanu, et al., 2001; Lee, 2006; Li & Baker, 2007; O'Connor, Buttrey, O'Dea, & Kennedy, 2011; Cui & Ye, 2015), and it has been reviewed and studied from a number of angles. Assessing and quantifying risk and safety in civil aviation has been the focus of many studies, and possible approaches for improving the safety of general aviation have been put forth (Janic, 2000; Li & Baker, 2007). In general, these researches can be divided into three main groups, as shown in Figure 1. Some researchers have studied aviation safety from a high-level managerial and administrative perspective (Cacciabue, Cassani, Licata, Oddone, & Ottomaniello, 2015; Oster Jr., Strong, & Zorn, 2013; Tamasi & Demichela, 2011; Davison, Ciavarelli, Cohen, Fischer, & Slovic, 2001; Netjasov & Janic, 2008). For instance, Netjasov & Janic (2008) describe four risk categories: (1) risk to an individual, (2) statistical risk that an accident will occur, (3) predicted risk, and (4) perceived risk. They also review different modelling methods of civil aviation risk and safety and divide these into four groups: (1) causal, (2) collision risk, (3) human-factor error, and (4) third-party risk.

Figure 1 - Summary of Research on Civil Aviation Safety



The second group of research includes the application of risk assessment methods in certain technical fields. Researchers in this group have investigated specific technical domains of aviation risks, such as airport properties; airplane systems control; aviation security screening; human factors, including pilot and air traffic controller; environmental impacts; and others. Airport-runway properties and their effects on aviation safety have been studied by researchers such as Waldron and Ford (2014), who investigated the airport runway's role in potential collisions and analysed how potential hazardous interactions can vary among airports. In a related vein, Galle et al. (2010) have examined runway incursions as a precursor to aviation accidents.

Another topic in this group is passenger security screening and how it affects aviation safety risks. Nikolaev, Lee, and Jacobson (2012) have studied the problem of multistage, sequential passenger screening with respect to passengers' risk levels. Mook and Scheinman (2011) have investigated risk-based screening systems to increase flight safety, while Stewart and Mueller (2013) introduced a method for risk-reduction estimation in commercial passenger airliners to prevent the aircraft from being hijacked.

Human error as a determining factor in aviation fatalities has also been studied in the second group. Nelson (1997) states that more than 50% of accidents and incidents in commercial aviation are caused by human error, and proposes a structured method to identify and correct potential human errors in aviation operations. Shyur (2008) has developed an analytical method to quantify aviation risks caused by human error, while Naranji, Mazzuchi, and Sarkani (2015) use augmented cognition and automated systems to reduce pilot error. Jin, Sun, and Kong (2010) examine the relationship between team situation awareness (SA) and information sharing, and propose a method to reduce human error. The authors also compare pilot SA and air traffic controller (ATC) requirements. Wei et al. (2012) have studied the main factors that influence human error in the cockpit, and developed a dynamic model for their prediction and evaluation. Human factors have also been studied from another perspective, which is flight-crew scheduling and the airline dispatcher's role in flight management. For instance, Graves et al. (1993) developed a new crew-scheduling system to reduce costs. The main concerns in such studies have been reducing costs, minimising flight delays, and optimising flight routing (Graves, McBride, Gershkoff, Anderson, & Mahidhara, 1993; Mercier & Soumis, 2007; Weide, Ryan, & Ehrgott, 2010; Nikulin & Drexl, 2010).

The third category includes studies that use mathematical and statistical models of civil aviation risks. Since this category is the most relevant to our research, we will discuss these in greater detail. Researchers have used a variety of mathematical tools to extract meaningful patterns from aviation safety databases. Some of the newer techniques, such as fuzzy logic, were applied by Lee (2006) to develop a quantitative model to assess aviation safety risk factors. The factors included in the model are evaluated based on their detectability, probability, criticality, etc. Other researchers have tried to capture patterns in the occurrence of accidents using more rigorous methods. Wang and Gao (2013) analysed the relationship between flight delays and aviation safety risk, and propose an approach based on Bayesian networks to model safety risk assessment. Another Bayesian-based model for avionic risk assessment was developed by Brooker (2011).

Causal methods can also be included in the third group; they are used to better determine how factors that affect the level of risk can be employed to evaluate overall risk (Netjasov & Janic, 2008). After each accident or incident, a causal report is prepared by related agencies in which they identify causal factors (Luxhøj & Coit, 2006). Janic (2000) classifies causal factors based on whether they are known or unknown and avoidable or unavoidable, and further differentiates causal factors based on accident type—i.e., whether they can be attributed to human error, mechanical failure, hazardous weather, sabotage, or military operations.

Spouge (2004) further discusses the benefits of causal analysis, and argues that safety managers and policy makers must understand the causes of accidents and evaluate the benefits of different intervention policies before selecting measures for risk reduction. Shyur, Keng, and Huang (2012) have developed an analytical model to analyse potential aviation events using both accident and performance measures; they employ an extended hazard-regression method to incorporate multiple safety performance indicators to assess the probability of aviation events. Their model may not be suitable for estimating absolute event probability, but it is valuable for understanding the structure of air events.

Common to these studies is the considerable emphasis placed on the use of different approaches to study flight accidents and incidents. These, in turn, funnel into data and prediction modelling. Underpinning these models are the data incorporated from aviation safety databases maintained by the Federal Aviation Administration (FAA), the National Transportation Safety Board (NTSB), and others, that have been used to model novel approaches to assess risk, capture patterns, and construct prediction models. Modelling the factors involved in aviation accidents/incidents has been at the core of these researches, which have focused on managing flight risk and increasing flight safety. The sheer range and diversity of these factors, however, significantly increases the difficulty of determining how each factor contributes to an event. Christopher and Balamurugan (2013) use data-mining approaches to predict aircraft accidents; they draw on the NTSB's aviation accident database, which does not include data on factors related to the pilot or weather. Because these variables offer vital insight into the causes of fatal aircraft accidents and improve data analysis, we have incorporated these factors in our database and will discuss them in detail in the following sections.

Nazeri et al. (2008) used a method called 'contrast-set mining of accidents and incidents' to interpret the relationship between those two and propose a model for accident-risk assessment. They found it difficult to identify a pattern in accidents, however, given the rarity of their occurrence—an observation well documented by Janic (2000), who highlights the difficulty in accurately locating, explaining, and managing overall aviation safety due to the scarcity of events. In turn, the former research favors incidents as the predominant tool in predicting the probability of an accident.

Though holistic in addressing all readily quantifiable data from either the FAA or NTSB databases, other factors that may have a significant impact on the analysis of risk are not included in these databases. Such factors are available, however, in NTSB Probable Cause Reports (PCRs). Capturing these factors entails close review of individual PCRs and translating relevant data points. Analyses that incorporate these factors would add robustness to already rigorous prior research and allow the consideration of additional factors. Nazeri et al. (2008) alludes to several such factors and notes, for example, the importance of an event's severity, phase of flight, and type of aircraft that, though unavailable in public databases, would significantly enhance the value of the information gained from the analysis.

Measuring how each factor affects an event—either individually or in combination—would offer researchers and decision-makers a deeper understanding of aviation events and, potentially, improve protocols and policies. It is worth mentioning that mathematical explanations of factual observations in aviation safety are also of great value. For instance, although the role of the pilot in flight safety seems obvious from an empirical point of view, one can only study the effect of pilot contributions in combination with other factors by using quantitative indices.

Differentiating and accentuating factors that have greater impact on events would save time, money, and human resources—and, ultimately, increase flight safety and efficiency. Therefore, investigating the relations between these factors—and specifically as dependent and independent variables using multivariate correlation analysis—is the main focus of this paper.

This study aims to examine how correlations between flight variables and incident/accident variables are affected by different factors. This emphasis on correlative analysis is intended to incorporate the aforementioned factors and demonstrate the approach's ability to yield highly specific results. Unlike researchers who have addressed the problem qualitatively, such as Nazeri et al. (2008), our goal is to first enlarge the aviation safety database by adding factors and values and then approach the problem quantitatively. This will not only yield qualitative results, but will also enable us to apply our findings to more advanced mathematical modelling that could be used by a variety of aviation personnel, such as flight dispatchers and crew schedulers, to optimise flight risk. For example, a flight dispatcher using the model could assess the risks imposed by weather on a specific flight against the risks imposed by pilots (i.e., the combined risks of the pilot and co-pilot) and plan the flight accordingly. The crew scheduler, in turn, could use the pilot variables to minimise risk by selecting the optimal combination of pilot and co-pilot.

The paper is organised as follows: Section 2 discusses how the current study's data were obtained, and how the raw public database was improved to allow for subsequent analysis. The section concludes by introducing dependent variables (DVs) and independent variables (IVs). Section 3 introduces the multivariate statistical analysis used, and Section 4 presents the results of our analytical method and discusses the significance of our findings. In Section 5 we present our conclusions and discuss avenues for future research.

2. DATA

To obtain meaningful results, we first required a comprehensive and reliable database. The second requirement was to define reasonable factors, including dependent and independent variables, and the third requirement was a statistical tool capable of measuring correlations between the variables. Careful selection of variables was crucial for our analysis. Criteria for data selection and methods for data pre-processing, variable selection, and grouping are described below. After building the database, a multivariate statistical method will be introduced and applied to reveal correlations among variables and identify the most influential.

Data Selection

The raw database for this research was obtained from the NTSB's database, which contains accident reports from 1962 to the present. Generally, a preliminary report is available online shortly after an accident occurs. As the NTSB investigation progresses, more data are added; upon completion of the investigation, the preliminary report is replaced by a final description of the accident and its probable cause (NTSB, 2014).

For a database to be downloaded, one must specify certain information and submit a relevant query. Preparing a database for retrieval often requires the provision of time intervals, locations, and the type of aircraft involved. The raw database used in this research was chosen from 10 different queries on the main NTSB repository; only accidents with published PCRs were considered. Table 1 shows the details of the query selected for the study, based on the data's relevance, functionality, and feasibility; data from other queries were either too cumbersome or too insignificant. The query selected includes 508 events, which comprise a sizable statistical population for data preparation.

Table 1. Selected Query Details

Query time interval	01/01/2003 to 12/31/2013
Location	USA
Aircraft category	Airplane

Operation type	Part 121: Air Carrier
Investigation type	Accident/Incident
Report Status	Probable Cause

In addition to the information provided in a downloadable spreadsheet, the PCR for each event (accident or incident) is available as a PDF and is more detailed than the information in the raw database.

The raw database was obtained and all corresponding PCRs downloaded. The database consisted of rows and columns in which rows correspond to events and columns to variables/factors. The database and PCR reports formed the basis for the process of data preparation and database enhancement.

Data Preparation

As mentioned above, the raw database retrieved from the NTSB lacked information pertinent to our study aims. We incorporated additional information as follows:
a) Grouping: Though public, the NTSB database is essentially intended for internal use; therefore, significant effort is required to prepare the database to perform statistical analysis. The first step was to group relevant factors into specific categories and reorder the variables' columns. For the purposes of this study, independent variables involving accidents/incidents were categorised according to type. Pilot information was not included in the original database, but because values were retrieved from PCRs in the next step and added to the database, a category was created for pilot information. Independent variables were divided into five categories:

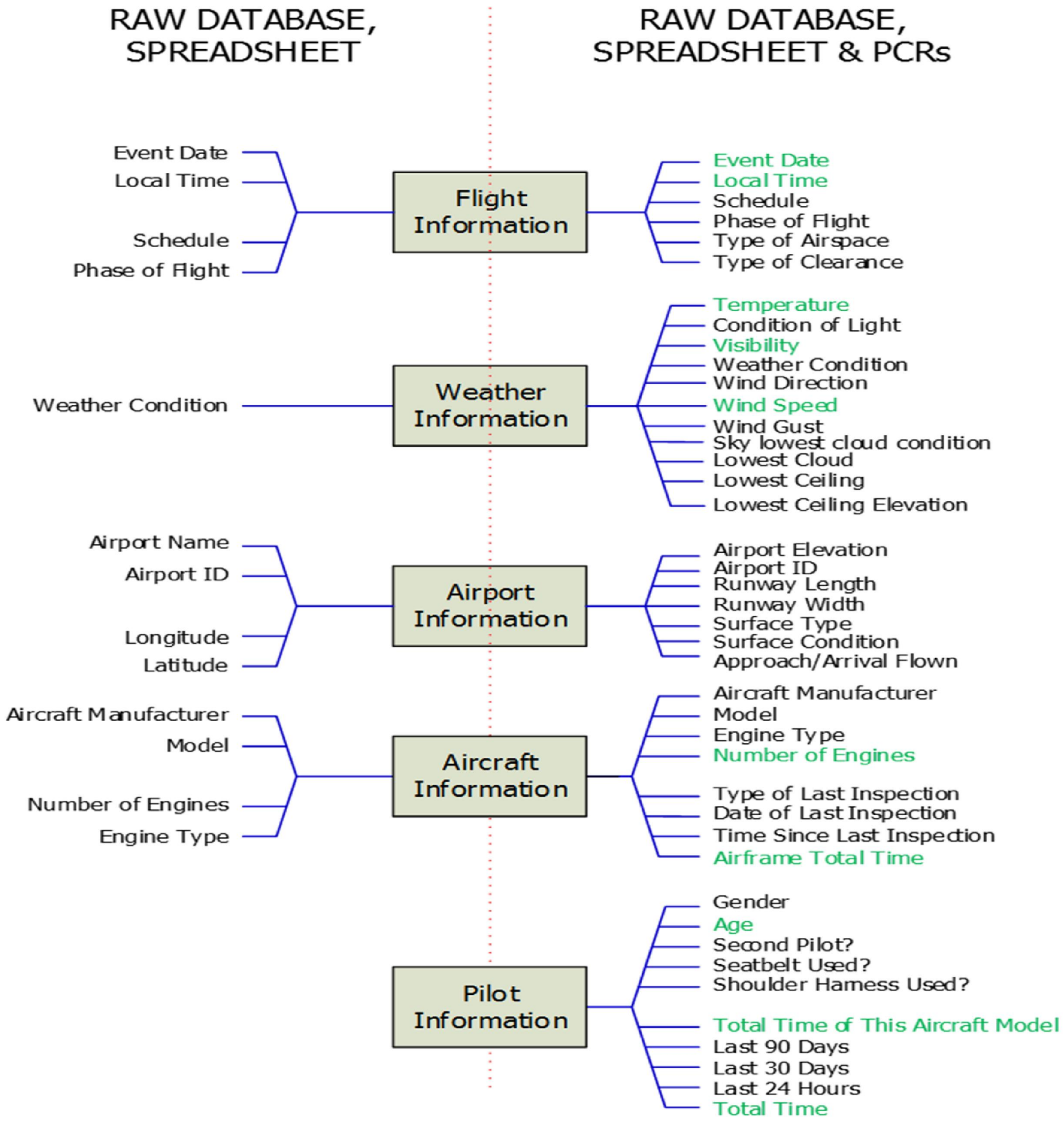
- Flight information
- Weather information
- Airport information
- Aircraft information
- Pilot information

We selected three dependent variables, which concern the magnitude of the event:

- Event type (accident or incident)
- Severity of injuries/number of fatalities
- Level of damage to aircraft

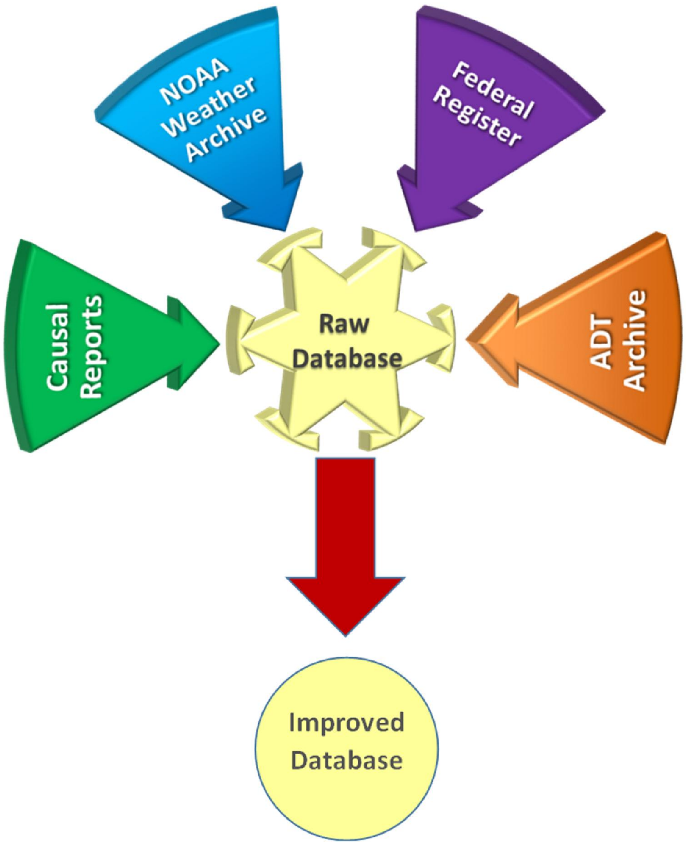
b) New Variables: The study includes several critical variables, such as pilot information, that are not provided in the raw NTSB database but are present, either explicitly or implicitly, in the detailed Probable Cause Reports (PCRs). These variables were selected based on advice from experts in the FAA and National Oceanic and Atmospheric Administration (NOAA). Once the variables had been chosen, individual PCRs were carefully examined to incorporate the new data into a more comprehensive database. Figure 2 shows the details of factors from the raw database and others that were collected from narrative PCRs. Data shown in green are those used in the final analysis, which will be discussed shortly.

Figure 2. Database Improvement using PCRs



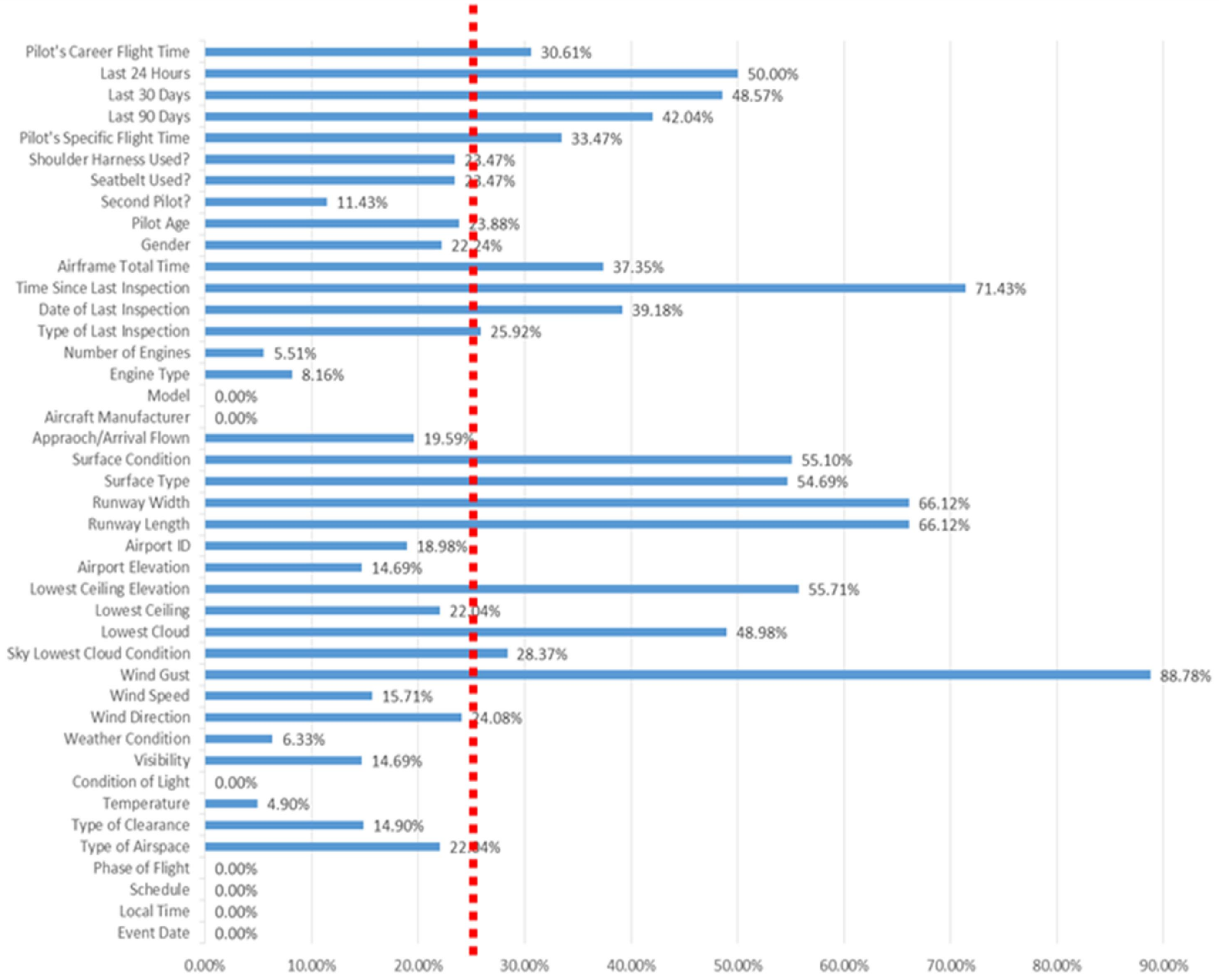
c) Data from Additional Sources: Grouping and including new variables expanded the database. In some instances, however, data for new variables—such as temperature, wind speed, visibility, airspace type, and airport elevation—were missing from either the raw databases or the PCRs. To acquire this information, we consulted sources other than the NTSB, such as the NOAA database for weather information, the average daily temperature (ADT) database of the University of Dayton, and the Federal Register for airport information. These external sources filled critical gaps in the raw database. Database improvement efforts are depicted in Figure 3. In some cases, the flight phase was not explicitly stated in the report, but was implicit in the narrative. In such cases, we based our judgment of the flight phase on the PCR's narrative.

Figure 3. Database Improvement Using Additional Sources



d) Database Cleaning: Given the sheer number of events under consideration and the range of variables, it was not possible to construct an exhaustive database. To ensure that the data collected would be relevant, we removed factors that were irrelevant or insufficiently significant (column cleaning). Likewise, events that were insufficiently significant or missing too many variables were removed (row cleaning). These steps were performed only after filling in as many gaps in the database as possible. The minimum acceptance threshold for variables was 25%—i.e., variables that were missing values for more than 25% of events were excluded. Percentages of missing values for each variable are shown in Figure 4; the red line represents the 25% threshold. Some variables were removed because they were almost uniformly constant—for example, the 'Shoulder harness used?' variable was either 'yes' or left blank in the PCR. There were also instances in which it was not possible to quantify value—for example, Airport ID and Type of Airspace are not quantifiable. To reveal their effects in the data analysis, however, they were included in the clustering phase, which will be discussed later.

Figure 4. Percentage of Missing Values for Independent Variables



In addition to the above, the date of the event (in the form of MM/DD) and the local time of occurrence (in the form of HH:MM) were normalised using the following formulas:

Date = (MM*30+DD)/365

Time = (HH+ (MM/60))/24

When it was necessary to convert qualitative data into quantitative data, we made logical assumptions. For example, wind speeds that were reported as 'calm' were assigned a numerical value of 0.5 mph. The database was now ready to perform statistical analyses, and independent and dependent variables had been finalised. Table 2 shows the resulting IVs and DVs, with information about type, range, and possible values for each variable.

Table 2. IVs and DVs for Statistical Analysis

	Independent Variables	Type and Possible Values	Unit
1	Event date	Normalised number between 0 an	N/A
2	Event time	Normalised number between 0 an	N/A
3	Phase of Fight	Standing, Taxi, Take Off, Climb, Descent, Approach, Landing	N/A
4	Temperature	Continuous values	Centigrade
5	Visibility	Continuous values	Statute Miles
6	Wind Speed	Continuous values	MPH
7	Number of Engines	Discrete values	N/A
8	Airframe Total	Continuous values	Hour
9	Age of pilot-in-command	Discrete values	Year
10	Pilot's Career Flight Time	Continuous values	Hour
11	Pilot's Specific Flight Tim accident/incident model)	Continuous values	Hour

	Dependent Variables (	Type and Possible Values	Unit
1	Event Type	Binary: Accident (2) or Incident (1)	N/A
2	Injury Severity	Discrete values: Number of fatalities	N/A
3	Level of aircraft damage	None (0) , Minor (1), Substanti Destroyed (3)	N/A

3. MULTIVARIABLE STATISTICAL ANALYSIS

To evaluate the effect of different IVs on the correlation between two sets of variables, a multivariate statistical analysis tool was necessary. In multivariate statistics, multivariate regression analysis is employed to investigate the relationship between a single DV and multiple IVs (Hair et al. 2010). In cases in which both dependent and independent variables are multivariate, the canonical correlation analysis (CCA) can be used to model the linear relationship between multiple DVs and multiple IVs (Borga 2001, Hardoon et al. 2004).

CCA and its Application

Prior research has demonstrated the uses and value of the CCA method to predict multiple DVs from multiple IVs (Bonner & Liu 2005, Singh et al. 2013, Singh et al. 2012). The aim with CCA is to identify and quantify the interrelations between a p-dimensional variable *X* and a q-dimensional variable *Y* (Dehon et al. 2000). The analysis looks for linear combinations of the original variables**, $\mathbf{a}^T\mathbf{X}$** and **$\mathbf{b}^T\mathbf{Y}$**, that have maximal correlation.

In mathematical terms, the CCA selects vectors $\alpha \in \mathrm{R}^p$ and $\beta \in \mathrm{R}^q$ such that:

$$(\alpha, \beta) = \mathrm{argmax}_{a,b}\,|\mathrm{Corr}(a^T X, b\,^T Y)|$$

The selected univariate variables, $U = X.\alpha$ and $V = Y.\beta$, are referred to as *canonical variates*. The number of pairs of *canonical variates* is equal to the minimum of *p* and *q.* Each pair of canonical variates interprets the relationship in a given way. The CCA method captures the highest correlation between linear combinations of IVs and linear combinations of DVs. The most significant pairs are those with the highest correlations (Nourzad & Pradhan, 2015). The single variables that represent *X*-values and *Y*-values, respectively, are created using the formulas below:

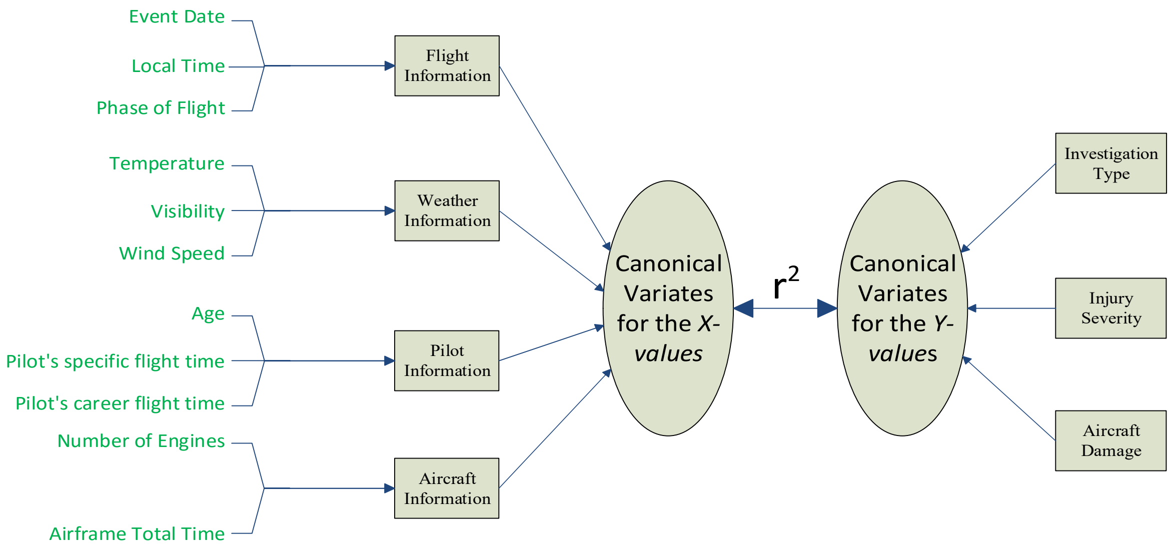
$$U = a_1.X_1 + a_2.X_2 + \cdots + a_p.X_p$$

$$V = b_1.Y_1 + b_2.Y_2 + \cdots + b_p.Y_p$$

We developed an approach to measure the correlation between DVs and IVs for flight accidents/incidents using the CCA method. MATLAB statistical toolbox functions (*canoncorr*) were used to run CCA. The first canonical correlation resulting from the MATLAB function is *the maximum correlation coefficient* between *U and V* for all *U* and *V* (Nourzad & Pradhan, 2015). The model's effectiveness depends on the goodness of fit of the captured linear relationships. The highest r-squared value (a measure of goodness of fit) corresponds to the most effective model for capturing relationships between *X*-values and *Y*-values. The main aim was to determine whether two sets of variables are related and, if so, how different variables affect the r-squared values.

As stated in the previous section, we selected *p=11* IVs and *q=3* DVs (accident-magnitude attributes) and used them to create canonical variates. The pairs with the highest r-squared values have the strongest correlations. Figure 5 depicts our model, in which the r-squared value will be measured and monitored depending on the change in the number of variables employed.

Figure 5 - Research Model



After database pre-processing, the first CCA run did not yield promising results. When all IVs were included, the r-squared value was 0.36, which signifies a weak correlation. We then performed clustering, which is a common approach in data analysis, to determine whether better results could be achieved without losing the selected IVs. Clustering is different from factors analysis; Cluster analysis tries to group cases/events that are more similar to each other than to other types of cases whereas factors analysis attempts to group features. Figure 6 is a generic illustration of how clustering can obtain stronger results from multivariate analysis.

To select the best variable to cluster, four variables capable of being clustered were chosen: Phase of Flight, Weather Condition, Flight Schedule, and Type of Clearance. Data clustering was then performed on each variable, and the resulting r-squared values compared. As shown in Figure 7, clustering based on Phase of Flight yielded the highest r-squared values.

Figure 6 - Data Clustering

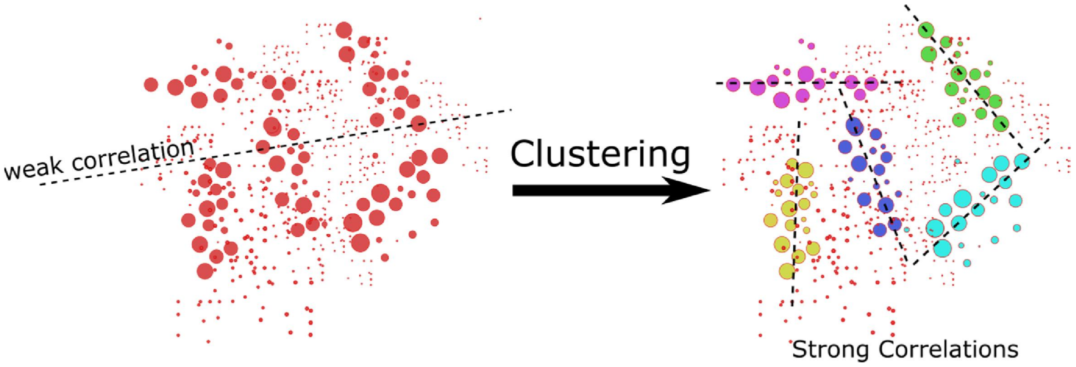
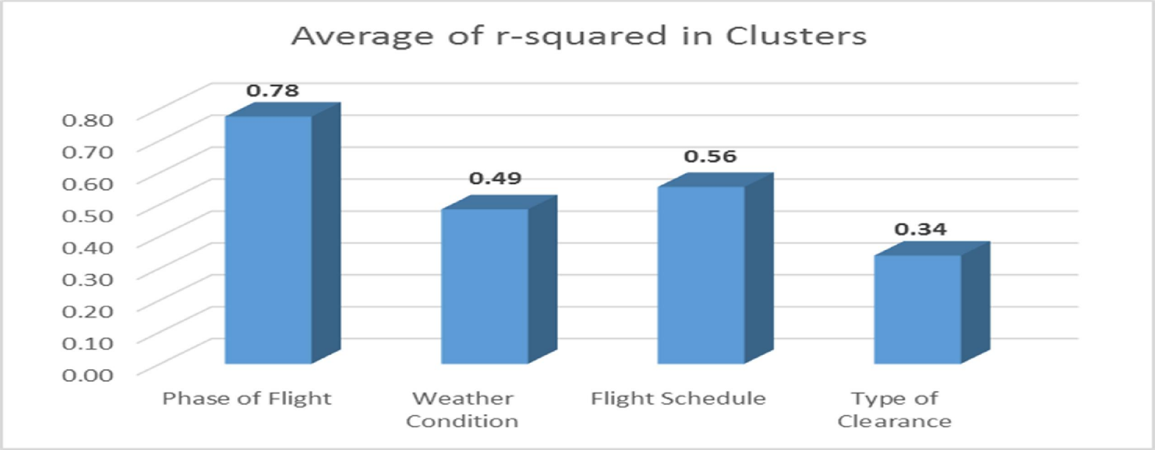


Figure 7 - Clustering Alternatives

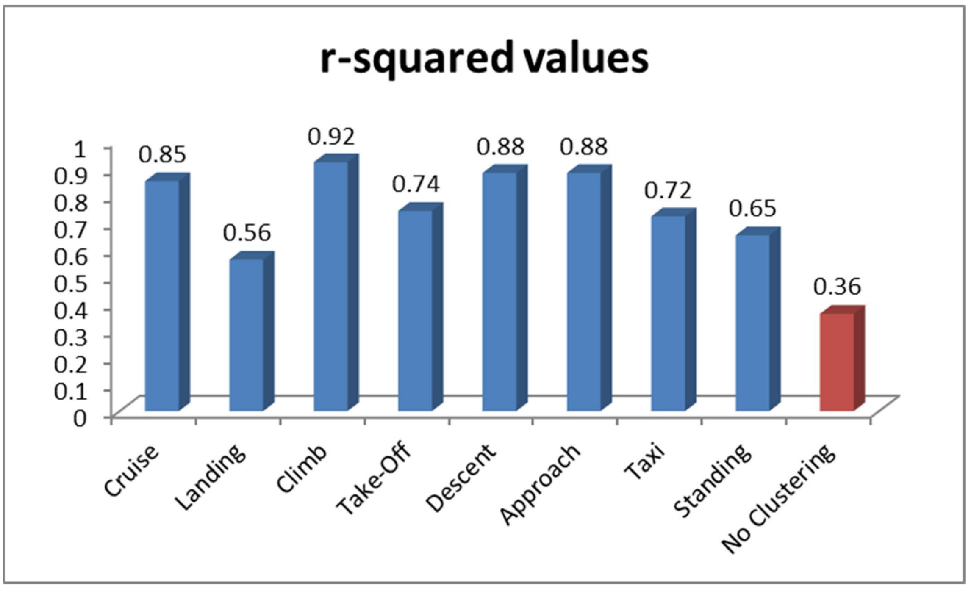


As explained in Section 1, Nazeri et al. (2008) recommended that future studies include the flight phase in which the accident/incident occurred. Together with results using other variables (Figure 7), this led us to select Phase of Flight as the variable for clustering. Eight flight phases were used as clusters for the database. To assess the effect of different variables on the correlation between canonical variates, CCA was performed multiple times on each cluster. Method details and results of the analysis are presented and discussed in the next section.

4. RESULTS AND DISCUSSION

As mentioned earlier, the first CCA run on the entire database did not yield fruitful results, since the r-squared value showed a weak correlation. Following clustering, the correlations were strengthened significantly. Clustering was based on Phase of Flight IV, which lent further relevance to nonnumerical values. The r-squared values for all eight phases, with and without clustering, are shown in Figure 8.

Figure 8 - Significance of Clustering



To investigate the effect of different variables on the r-squared values for each cluster, the CCA statistical test was run six times with different variables. The first run included all IVs. Successive runs were performed by excluding one group of IVs at a time while recording the resulting changes. Consider, for example, the Cruise cluster, which includes all events that occurred during that phase. The first run obtained 0.85 for the highest r-squared value between canonical variates. The second run included all IVs except Weather Information. The resulting r-squared value was 0.84, showing a minimal decrease in correlation. The third run was performed including all IVs except Pilot Information. The resulting r-squared value was 0.56, showing a significant drop in correlation (34%). This supports the claim that the effect of pilot-associated information is much more significant than weather information in the investigation of correlations between different factors of flight events. Remaining runs were performed in the same manner as the Cruise phase for the other seven Phases of Flight. Detailed results are shown in Table 3.

Table 3 - r-squared Values for Different Run of CCA

		Cruise	Landing	Climb	Take-Of	Descent	Approach	Taxi	Standing	All 8
1	All factors	0.85	0.56	0.92	0.74	0.88	0.88	0.72	0.65	0.36
2	Without Flight Info	0.82	0.52	0.89	0.68	0.84	0.78	0.60	0.56	0.35

3	Without Weather Info	0.84	0.47	0.81	0.67	0.86	0.83	0.65	0.52	0.33
4	Without pilot info	0.56	0.47	0.91	0.60	0.81	0.69	0.47	0.62	0.26
5	Without Aircraft Info	0.83	0.56	0.88	0.74	0.83	0.80	0.70	0.63	0.32

To gain a better understanding of the effect of different variables on goodness of fit, it is necessary to calculate the level of drop in r-squared values when each group is excluded from the analysis. Drops are calculated as percentages and shown in Table 4 and Figure 9. As shown in Table 3, in five out of eight flight phases, pilot-associated data played the most significant role in the correlation between DVs and IVs for accidents/incidents. This phenomenon was observed by removing pilot-associated variables and monitoring the changes in other variables. The highest drops are seen in the Taxi, Cruise, Approach, and Take-off phases.

Table 4 - Drop in r-squared Values in %

	Cruise	**Landing**	**Climb**	**Take-O**	**Descent**	**Approach**	**Taxi**	**Standing**	**All 8**
Drop in r^2 for Flight Info	4%	7%	3%	8%	5%	11%	17%	14%	3%
Drop in r^2 for Weather Info	1%	16%	12%	9%	2%	6%	10%	20%	8%

Drop in r^2 for Pilot Info	34%	16%	1%	19%	8%	22%	35%	5%	28%
Drop in r^2 for Aircraft Info	2%	0%	4%	0%	6%	9%	3%	3%	11%

Our findings further demonstrate that even without clustering, pilot information has the greatest effect of the IVs in all but two flight phases, Standing and Climb. The lower level of correlation in the Standing phase can be attributed to the pilot's low level of involvement; it is reasonable that other factors, such as weather information or airport-related factors, would be more influential, and this is corroborated by the results shown in Figure 9. In the case of the Climb phase, the discrepancy in correlation may be attributed to the low number of events recorded during this phase. The overall process of preparing the database, performing multivariate statistical tests, and obtaining results is illustrated in Figure 10.

Figure 9 - Drop in r-squared Values for Different Flight Phases

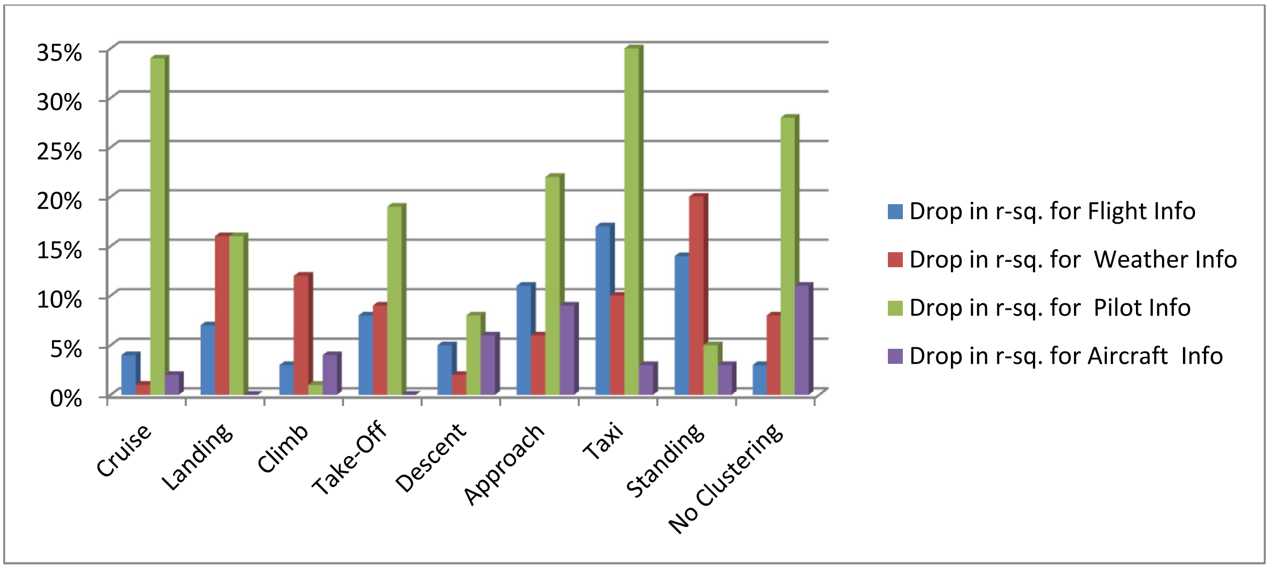
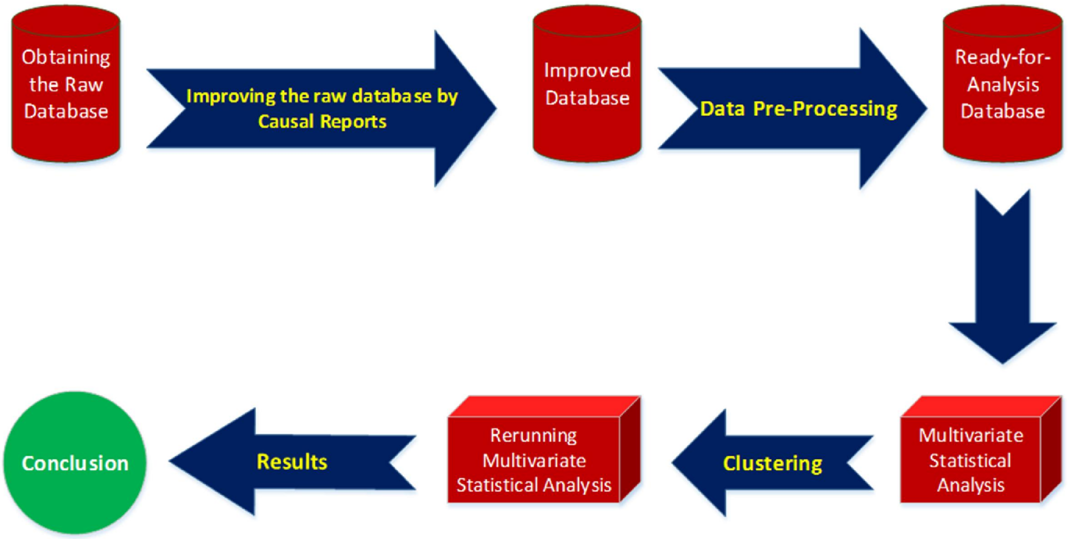


Figure 10 - An Overview of Research Methodology


Obtaining the Raw Database
Improving the raw database by Causal Reports
Improved Database
Data Pre-Processing
Ready-for-Analysis Database
Multivariate Statistical Analysis
Clustering
Rerunning Multivariate Statistical Analysis
Results
Conclusion

5. CONCLUSIONS AND FUTURE STUDIES

In this paper, CCA was used to analyse an enhanced aviation safety database to identify the effects of different variables on correlations between flight factors and event factors. The study's focal point was to identify and assess relevant factors in aviation events. Prior research with a similar aim has lacked a comprehensive database that incorporates not only raw information from the NTSB, but, as with this study, additional data from sources that are not immediately quantifiable (e.g., the NTSB's PCRs). Database enhancement was performed by studying all associated PCRs and retrieving new variables. The enhancement process included grouping, introducing new variables, obtaining data from additional sources, and database cleaning. Having said that, this research was limited to events happened in USA and mentioned in the NTSB main database. The next step was to determine whether the enhanced database would be suitable for CCA, with the goal of discovering the most influential factor among the IVs considered. Initial results were not promising, so a clustering method was proposed. Clustering based on Phase of Flight was selected after comparing clustering options. CCA was run six times in each cluster with different variables, based on the research model, to investigate the variables' effects on r-squared values between DVs and IVs.

Our findings statistically support the empirical observation that pilot-associated data, including age, career flight time, and experience with the aircraft model involved in the event, are the most effective factors in demonstrating a correlation between dependent and independent variables of aviation events. The second, third, and fourth most significant factors were variables associated with weather, flight time, and aircraft, respectively.

This research provides a framework for further inquiry and the construction of a predictive model using the more comprehensive database we have made available. Such a predictive model could be used by different stakeholders, such as risk managers, airline planners, crew schedulers, and dispatchers, to minimise flight risk and improve flight safety. These findings could be used to improve flight-crew scheduling and dispatching practices; consideration of these factors when selecting pilots and co-pilots could also reduce flight risk. Prior entering raw data in regular flight scheduling process, the above mentioned predictive model can be used to assess the combination of those factors and the level of risk they impose. This model can potentially tell schedulers that in specific weather conditions, *how* assigning low experience pilot will increase the risk of flight. This model can also be used to reduce the risks based on the known variables prior to flight. "Flight variable assessment" based on this model can be added into existing flight scheduling processes to measure the level of risks imposed by flight variable combination. For example, a pilot with more experience and higher variable values could be paired with a low-hours co-pilot with less experience to optimise flight risk and, possibly, lower cost. Likewise, if weather factors based on our findings were included in the crew- scheduling process, better results might be obtained. By evaluating the risks prior to flight, the dispatcher or flight-crew scheduler could modify and reroute the flight, if necessary, based on weather conditions and pilot variables.

CCA was applied in this research so it imposes its limitations and assumptions. Linear relationship assumed for all variables in each set and also between sets. Applying none-linear methods can improve results and contribute to findings of our study. Widening the events selection criteria and including other countries aviation events, can potentially improve the results. Our method is also adaptable for a wide range of research topics. Other analytic methods, such as neural network analysis or fuzzy logic, could be used to determine whether similar results can be obtained.

REFERENCES

- Bonner, A., & Liu, H. (2005). Canonical Correlation, an Approximation, and the Prediction of Protein Abundance. *Eighth Workshop on Mining Scientific and Engineering Datasets (MSD'05)*, (pp. 29-38). Newport Beach.

- Brooker, P. (2011). Experts, Bayesian Belief Networks, rare events and aviation risk estimates. *Safety Science, 49*, 1142-1155.
- Cacciabue, P. C., Cassani, M., Licata, V., Oddone, I., & Ottomaniello, A. (2015). A Practical Approach To Assess Risk In Aviation Domains For Safety Management Systems. *Cogn Tech Work, 17*, 249-267. doi:10.1007/s10111-014-0294-y
- Christopher, A., & Balamurugan, S. (2013). Data Mining Approaches for Aircraft Accidents Prediction: An Empirical Study on Turkey Airline. *IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013)*, (pp. 739-745).
- Cui, Q., & Ye, L. (2015). The Change Trend and Influencing Factors of Civil Aviation Safety Efficiency: The Case of Chinese Airline Companies. *Safety Science, 75*, 56-63.
- Davison, J., Ciavarelli, A., Cohen, M., Fischer, U., & Slovic, P. (2001). The Many Faces of Risk in Aviation Decision-Making, The Human Factors and Ergonomics Society *45th Annual Meeting,* 307 - 310.
- Dehon, C., Filzmoser, P., & Croux, C. (2000). Robust Methods for Canonical Correlation Analysis. *Data Analysis, Classification, and Related Methods*, 321-326.
- Galle, K. M., Ale Jr., J. C., Hossain, M. M., Moliterno, M. J., Rowell, M. K., Revenko, N. V., . . . Haimes, Y. Y. (2010). Risk-Based Airport Selection for Runway Safety Assessments Through the Development and Application of Systems-Driven Prioritization Methodologies. *2010 IEEE Systems and Information Engineering Design Symposium.* Charlottesville.
- Graves, G. W., McBride, R. D., Gershkoff, I., Anderson, D., & Mahidhara, D. (1993, June). Flight Crew Scheduling. *Management Science, 39*(6), 736 - 745.
- Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation, 16*, 2639-2664.
- Janic, M. (2000). An Assessment of Risk and Safety in Civil Aviation. *Journal of Air Transport Management, 6*, 43-50.

- Jin, H., Sun, R., & Kong, X. (2010). Enhance Team Situation Awareness by Sharing Information to avoid Human Errors in Aviation. *Third International Conference on Knowledge Discovery and Data Mining.*
- Lazarick, R. T. (1998). Airport Vulnerability Assessment - An Analytical Approach. *SPIE 3575, Enforcement and Security Technologies.*
- Lee, W.-K. (2006). Risk Assessment Modeling in Aviation Safety Management. *Journal of Air Transport Management, 12*(5), 267-273.
- Li, G., & Baker, S. P. (2007). Crash Risk in General Aviation. *Journal of American Medical Association, 297*(14), 1596-1598.
- Luxhøj, J. T., & Coit, D. W. (2006). Modeling Low Probability/High Consequence Events: An Aviation Safety Risk Model. *Proceedings of the 2006 Reliability and Maintainability Symposium (RAMS).* Newport Beach.
- Mercier, A., & Soumis, F. (2007). An integrated aircraft routing, crew scheduling and flight retiming model. *Computers & Operations Research, 34*, 2251 - 2265.
- Mook, D., & Scheinman, E. D. (2011). Risk Based Screening and Explosive Detection at the Passenger Screening Checkpoint. *2011 IEEE International Conference on Technologies for Homeland Security (HST)*, (pp. 98-103).
- Naranji, E., Sarkani, S., & Mazzuchi, T. (2015, June). Reducing Human/Pilot Errors in Aviation Using Augmented Cognition and Automation Systems in Aircraft Cockpit. *Transactions on Human-Computer Interaction, 7*(2), 71 - 96.
- Nazeri, Z., Barbara, D., Jong, K. D., Donohue, G., & Sherry, L. (2008). Contrast-Set Mining of Aircraft Accidents and Incidents. In *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects* (Vol. 5077, pp. 313-322). Springer Berlin Heidelberg. doi:10.1007/978-3-540-70720-2_24
- Nelson, W. R. (1997). Structured Methods for Identifying and Correcting Potential Human Errors in Aviation Operations. *1997 IEEE International Conference On Systems, Man, And Cybernetics. 4*, pp. 3132-3136. IEEE.

- Netjasov, F., & Janic, M. (2008). A review of research on risk and safety modelling in civil aviation. *Journal of Air Transport Management, 14*, 213-220.
- Nikolaev, A. G., Lee, A. J., & Jacobson, S. H. (2012). Optimal Aviation Security Screening Strategies With Dynamic Passenger Risk Updates. *IEEE Transactions on Intelligent Transportation Systems, 13*(1), 203-212. doi:10.1109/TITS.2011.2167230
- Nikulin, Y., & Drexl, A. (2010). Theoretical aspects of multicriteria flight gate scheduling: deterministic and fuzzy models. *Journal of Scheduling, 13*, 261 - 280. doi:10.1007/s10951-009-0112-1
- Nourzad, S. H., & Pradhan, A. (2015). Computational Modeling of Networked Infrastructures: A Macroscopic Multivariate Approach. *Journal of computing in civil engineering*.
- O'Connor, P., Buttrey, S. E., O'Dea, A., & Kennedy, Q. (2011). An Assessment of Safety Climate in U.S Naval Aviation. *the Human Factors and Ergonomics Society 55th Annual Meeting*, (pp. 1740-1744). doi:10.1177/1071181311551361
- Orasanu, J., Davison, J., Ciavarelli, A., Cohen, M., Fischer, U., & Slovic, P. (2001). The Many Faces of Risk in Aviation Decision Making. *the Human Factors and Ergonomics Society 45th Annual Meeting*, (pp. 307-310).
- Oster Jr., C. V., Strong, J. S., & Zorn, C. K. (2013). Analyzing Aviation Safety: Problems, Challenges, Opportunities. *Research in Transportation Economics, 43*, 148-164.
- Shyur, H.-J. (2008). A Quantitative Model for Aviation Safety Risk Assessment. *Computers & Industrial Engineering*, 34-44.
- Shyur, H.-J., Keng, H., I, I.-K., & Huang, C.-L. (2012). Using Extended Hazard Regression Model to Assess the Probability of Aviation Event. *Applied Mathematics and Computation, 218*, 10647-10655.
- Singh, A., Acharya, N., Mohanty, U. C., & Mishra, G. (2013). Performance of Multi Model Canonical Correlation Analysis (MMCCA) for Prediction of Indian Summer Monsoon Rainfall Using GCMs Output. *Comptes Rendus Geoscience, 345*, 62-72.

- Singh, A., Kulkarni, M. A., Mohanty, U. C., Kar, S. C., Robertson, A. W., & Mishra, G. (2012). Prediction of Indian Summer Monsoon Rainfall (ISMR) Using Canonical Correlation Analysis of Global Circulation Model Products. *Meteorological Applications, 19*(2), 179-188. doi:10.1002/met.1333
- Spouge, J. (2004). *A Demonstration Causal Model for Controlled Flight into Terrain.* London: Det Norske Veritas.
- Stewart, M. G., & Mueller, J. (2013). Aviation Security, Risk Assessment, and Risk Aversion for Public Decisionmaking. *Journal of Policy Analysis and Management, 32*(3), 615-633. doi:10.1002/pam.21704
- Tamasi, G., & Demichela, M. (2011). Risk Assessment Techniques for Civil Aviation Security. *Reliability EngineeringandSystemSafety, 96*, 892-899.
- Waldron, T. P., & Ford, A. T. (2014). Investigating the Causality of Potential Collisions on the Airport Surface. *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, (pp. 7C5-1-7C5-11). Colorado Springs. doi:10.1109/DASC.2014.6979520
- Wang, H., & Gao, J. (2013). Bayesian Network Assessment Method for Civil Aviation Safety Based on Flight Delays. *Mathematical Problems in Engineering*. Retrieved from http://dx.doi.org/10.1155/2013/594187
- Wei, Z., Zhuang, D., Wanyan, X., Wei, H., & Zhou, Y. (2012). Prediction and Analysis of the Human Errors in the Aircraft Cockpit. *2012 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012).* IEEE.
- Weide, O., Ryan, D., & Ehrgott, M. (2010). An iterative approach to robust and integrated aircraft routing and crew scheduling. *Computers & Operations Research, 37*, 833 - 844.