RISK PRECURSOR ANALYSIS: AN ASSESSMENT OF FLIGHT VARIABLES AFFECTING CIVIL AVIATION ACCIDENTS AND INCIDENTS

by **Mohammad Kashef**

B.S. in Civil Engineering, September 2007, Iran University of Science and Technology
M.S. in Construction Management, August 2011, Middle East Technical University

A Dissertation submitted to

the Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Doctor of Philosophy

August 31, 2016

Dissertation directed by

Thomas A. Mazzuchi
Professor of Operations Research and of Engineering Management
and
Shahram Sarkani
Professor of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Mohammad Kashef has passed the Final Examination for the degree of Doctor of Philosophy as of July 26, 2016. This is the final and approved form of the dissertation.

RISK PRECURSOR ANALYSIS: AN ASSESSMENT OF FLIGHT VARIABLES AFFECTING CIVIL AVIATION ACCIDENTS AND INCIDENTS

Mohammad Kashef

Dissertation Research Committee:

Shahram Sarkani, Professor of Engineering Management and Systems Engineering, Dissertation Co-Director

Thomas Mazzuchi, Professor of Engineering Management and Systems Engineering & of Decision Sciences, Dissertation Co-Director

Timothy Blackburn, Professorial Lecturer in Engineering Management and System Engineering, Committee Member

Bill Olson, Professorial Lecturer in Engineering Management and System Engineering, Committee Member

E. Lile Murphree, Professor Emeritus of Engineering Management and Systems Engineering, Committee Member

© Copyright 2016 by Mohammad Kashef, All rights reserved

DEDICATION

This work is dedicated to my beloved parents, whose support and prayers are always with me; my loving wife, who helps me overcome life's challenges; my lovely daughters and handsome son, who are always a relief after a tedious day of work.

ACKNOWLEDGEMENT

This research was accomplished with the support and help of different individuals whom I am honored and privileged to have worked alongside. I am deeply thankful to my PhD advisers, Dr. Shahram Sarkani and Dr. Thomas Mazzuchi who guided me through this research from its early stages to its final destination. Their ideas, knowledge, and expertise added much value to this work. Their critical views and the challenges they brought up helped a great deal in shaping the quality of this research.

I must also extend my appreciation to Dr. Hossein Nourzad, who initiated the idea for this research and shared his valuable experience in the modeling and data analysis phases.

I wish to express appreciation to Dr. Ehsan Naranji who supported me through this research with his constant encouragement and creative solutions to overcome various obstacles I faced during research.

I am speechless in thanking my dear parents who continuously and devotedly supported my effort and so tirelessly encouraged me throughout this process; and my wife who has always been there for me—without her I would never have achieved such success.

ABSTRACT

Risk Precursor Analysis: An Assessment Of Flight Variables Affecting Civil Aviation Accidents And Incidents

Flight safety has been an important topic for both academia and the industry. Aviation experts and authorities, as well as commercial airline administrators, have always been looking for improvements in the safety of flights. Researchers, on the other hand, have tried to model avionic fatalities and suggest improvements or upgrades in the flight systems to reduce flight risks. One perspective has been to look at available data from past accidents and incidents to capture and model the relationship between different factors involved in each event. Some important factors are not included in the databases maintained by authorities like the National Transportation Safety Board (NTSB). First, this research tries to improve existing database factors by adding more variables; second, it assesses the level of effect each group of independent variables has on the correlation between flight factors and accident/incident magnitude factors using a multivariate statistical analysis. This can be considered as accident/incident precursor analysis in aviation risk management. The study concludes that the pilot related factors are the most influential variable group on the correlation between Dependent Variables (DVs) and Independent Variables (IVs). The significance of these factors or group of factors helps researchers and policy makers to consider the more important ones in their decision-making processes. The research extends its contribution by building a descriptive

model using canonical analysis for different phases of flight. This model can help flight schedulers as a decision-supporting tool in their scheduling process.

TABLE OF CONTENTS

DEDICATION ..... **iv**

ACKNOWLEDGEMENT ..... **v**

ABSTRACT ..... **vi**

TABLE OF CONTENTS ..... **viii**

LIST OF FIGURES ..... **xi**

LIST OF TABLES ..... **xiii**

Chapter 1 – INTRODUCTION **1**

1.1 Problem Background 1

1.2 Aviation Safety Databases 3

1.2.1 FAA 3

1.2.2 National Transportation Safety Board (NTSB) Database 5

1.3 Research Overview 6

Chapter 2 – LITERATURE REVIEW **8**

2.1 Risks in Civil Aviation 8

2.2 Risk Measurement in Civil Aviation 10

2.3 Accident Causation Theories 12

2.3.1 The SHELL Model of Human Factors in Aviation Safety 12

2.3.2 Swiss Cheese Model on Aviation Safety 14

2.3.3 The Domino Effect Theory on Aviation Safety 16

2.4 Aviation Safety Studies 18
2.5 Statistical Methods 25
2.5.1 Introduction 25
2.5.2 Correlation 27
2.5.3 Multivariable Statistical Analysis 30
2.5.3.1 Advantages of Multivariate Statistics 31
2.5.3.2 Classification of Multivariate Techniques 32
2.5.4 Canonical Analysis 35
2.5.4.1 Canonical Correlation Analysis 35
2.5.4.2 CCA Terminology 37
2.5.4.3 Advantages of CCA 38
2.5.4.4 Appropriate Utilization of CCA 39
2.5.4.5 CCA Assumptions 41
2.5.5 Theoretical Foundations 42
2.5.5.1 Multiple Correlation Review 42
2.5.5.2 Mathematics of CCA 43
2.5.5.3 Canonical Correlations Properties 47
2.5.5.4 Canonical Correlation Applications 47
Chapter 3 – RESEARCH APPROACH 49
3.1 Research Hypothesis and Questions 50
3.2 Data 51
3.3 Data Selection 51
3.4 Data Preparation 54

3.5 Research Methodology 62

3.6 CCA and its Application 65

Chapter 4 – RESEARCH ANALYSES 72

4.1 Results 72

4.2 Training Canonical Model – One Phase of Flight 76

4.3 Model Verification – One Phase of Flight 80

4.4 Training and Verification of Canonical Model: Other Phases of Flight 81

4.5 Training Canonical Model – All Phases of Flight 83

4.6 Model Verification – All Phases of Flight 84

4.7 Discussion 85

Chapter 5 CONCLUSION 89

REFERENCES 94

Appendix A – Sample of PCR 99

LIST OF FIGURES

Figure 2-1 Air transport subsystems 8

Figure 2-2 SHEL Model 13

Figure 2-3 Swiss Cheese Model, (Modified from (Reason, 1990)) 15

Figure 2-4 Domino Model of Accident Causation (Modified from Heinrich, 1931) 17

Figure 2-5 Summary of Research on Civil Aviation Safety 19

Figure 2-6 Data Structure for Different Multivariate Techniques, modified from (McGarigal, Cushman, & Stafford, 2000) 34

Figure 2-7 Conceptual model for canonical correlation, modified from (Sherry & Henson, 2005) 41

Figure 3-1 Database Improvement Using PCRs 57

Figure 3-2 Database Improvement Using Additional Sources 59

Figure 3-3 Percentage of Missing Values for Independent Variables 60

Figure 3-4 An Overview of Research Methodology 63

Figure 3-5 Research Model 67

Figure 3-6 Data Clustering 68

Figure 3-7 Correlation between DVs and IVs in Each Phase of Flight 69

Figure 3-8 Correlation between DVs and IVs in Each Type of Flight 69

Figure 3-9 Correlation between DVs and IVs in Different Weather Conditions 69

Figure 3-10 Correlation between DVs and IVs in Different Clearance Conditions 70

Figure 3-11 Average of r-squared Value for Different Clusters 70

Figure 4-1 Significance of Clustering 72

Figure 4-2 Drop in r-squared Values for Different Flight Phases 76

Figure 4-3 Canonical Variates 1 and 2 79

Figure 4-4 Percentages of Training and Verifying Data Points 82

Figure 4-5 Combination of Testing Events 84

LIST OF TABLES

Table 2-1 Points of interest in the historical backdrop of Correlation and Regression (Rodgers & Nicewander, 1988) 28

Table 2-2 Multivariate techniques classification based on data structure, modified from (McGarigal, Cushman, & Stafford, 2000) 32

Table 2-3 Pairs of Canonical Variates 45

Table 3-1 Details of different queries on NTSB database 53

Table 3-2 Selected Query Details 54

Table 3-3 IVs and DVs for Statistical Analysis 61

Table 4-1 r-squared Values for Different Run of CCA 74

Table 4-2 Drop in r-squared Values in % 75

Table 4-3 Summary of Average Error in Canonical Model 83

Table 5-1 Research Methodology Summary 90

Table 5-2 Group Factors in the Order of Their Significance 91

Chapter 1 – INTRODUCTION

This chapter presents background information on aviation risk management and an overview of the research process.

1.1 Problem Background

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 associated costs, covers a broad range of disciplines ranging from risk management to flight-crew scheduling.

Despite major technological developments in the field, fatal accidents—often with high numbers of casualties—occur with alarming frequency. Recent crashes, in particular Malaysian Airlines 370 (on March 8, 2014, this flight disappeared carrying 239 people), Air Asia 8501 (lost in the Java Sea during storms, killing the 155 passengers it carried (Secret Flying, 2016)), and Germanwings 9525 (which was deliberately crashed by the co-pilot, killing all 150 people on board), highlight the critical importance of flight safety. One of the major reasons for accidents in the field of aviation, beyond system failures, is

the human error (Netjasov & Janic, 2008). These errors include traffic controller error, pilot and co-pilot error, safety inspector crew error, and so on. There are, however, other factors that contribute to accidents and their severity: weather condition, airport condition, or aircraft age being a few.

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 aircraft accident likewise is an event that results in the demise of individuals, injury and/or extensive damage to the aircraft. Boeing defines an aircraft accident as an event connected to the operation of an aircraft, during the time people with intent to fly are onboard the plane and before they disembark, in which (Boeing, 2015):

- Considerable damage occurs to the aircraft.
- The airplane goes missing or cannot be accessed.
    - Missing is defined as when the official search for the vessel has ceased and there is no trace of it to be found.
- The airplane causes fatality or grave injury, (i.e. there is direct exposure to jet blast, or direct contact with the aircraft).

For example, two planes that crash into one another on a runway would be considered an accident (in this case, most likely due to substantial aircraft damage). Two planes preparing for takeoff without required clearance of each other, meanwhile, would be documented as an incident.

In general, causes of accidents focusing on the correlation between different factors and the severity of events have not yet been studied in detail. This study has tried to build

a thorough database using existing flight events databases and statistical tools to analyze the correlation between factors involved in an event. It will present a model which assesses which factor(s) have more significant effects on event severity and how that information will help reduce flight risks prior to scheduling and crew assignment.

1.2 Aviation Safety Databases

Because of the importance of flight safety multiple agencies document aviation events. This provides various data for researchers to study towards the increasing of flight safety and the reduction of casualties. In the U.S., the Federal Aviation Administration (FAA), the National Transportation Safety Board (NTSB), and National Aeronautics and Space Administration (NASA) are the main bodies that prepare such reports and maintain accident/incident databases. The following are some of the U.S. civil government aviation safety databases (Accident/Incident Data Analysis Database Summaries - Vol. I – MARCH 1989):

1.2.1 FAA

Several databases are being maintained by FAA that may offer information pertinent to human factors research. Some of them are categorized as subsystems of a broader data base system. The systems examined and the pertinent subsystems are listed below

1.2.1.1 FAA Aviation Safety Analysis System (ASAS)

Established in 1982, the ASAS aim was supporting the Office of Aviation Standards. The ASAS is the FAA program for collecting and organizing various types of aviation safety data into an

integrated system and for using sophisticated analysis methods to analyze the data. Under FAA ASAS system, there are 4 sub-systems which are summarized below:

1.2.1.2 FAA Accident/incident Data Subsystem (AIDS) Database

1.2.1.3 FAA Enforcement Information Subsystem (EIS) Database

1.2.1.4 FAA Service Difficulty Reporting Subsystem (SDRS) Database

1.2.1.5 FAA Comprehensive Airmen Information Subsystem (CAIS) Database

1.2.1.6 FAA National Airspace Incident Monitoring System (NAIMS)

The NAIMS was established in 1985 to provide the FAA with a better basis for monitoring events that occurred in the National Airspace System that could impact aviation safety. NAIMS is a family of automated databases that consists of Pilot Deviations (PDs), Operational Errors (OEs), and Near-Mid-Air Collisions (NMACs).

The data are primarily statistical and contain limited information as to why human error has occurred. Under this database, 3 sub-systems are maintained which are categorized as below:

- Pilot Deviation (PO)

- Operational Error (OE)
- Near-Mid-Air Collision (NMAC)

1.2.1.7 FAA Air Carrier Accident Data System (ACADS) Database

The purpose of the database is to assist the Office of Accident Investigation in tracking all air carrier and scheduled commuter aircraft accidents.

1.2.1.8 FAA Aviation Safety Reporting system (ASRS) Database

The ASRS supplies data to the FAA and those involved with aviation to help the FAA reach toward the removal of unsafe conditions and to prevent avoidable accidents (Murphy & Levendoski, 1989).

1.2.2 National Transportation Safety Board (NTSB) Database

The NTSB is tasked by Congress to explore transportation accidents to determine cause(s) or probable cause(s) and to make safety recommendations aimed at preventing future accidents. To assist in carrying out this task, the NTSB established a computerized database. The NTSB's aviation accident database has maintained data on aircraft accidents and incidents throughout the United States, including territories, and over international waters since the early 1960's. The NTSB defines the most likely reason for each event and presents safety recommendations in order to help prevent other accidents with similar circumstances.

> The database can be retrieved in different formats. One of the formats is categorized based on type of event, including tables, like the inclusion of a narrative field that identifies the probable cause of each event occurrence (Naranji, Sarkani, & Mazzuchi, 2015).

These databases maintain valuable but raw data from each crash or aviation incident. This provides a platform for researchers to analyze causes of accidents and apply statistical methods to interpret data and provide predictive models.

1.3 Research Overview

As mentioned earlier, safety of air transportation has been a major concern for different governmental agencies as well as academia.

In general, previous studies conducted by other researchers regarding the subject of civil aviation safety are grouped into 3 categories: 1) Policies, strategies, and managerial studies. These studies look at safety aviation from a high-level managerial perspective. 2) Application Oriented Studies. In these studies certain technical fields are considered such as airport properties, airplane systems control, human factors, etc. 3) Mathematical and Statistical Modeling Approaches. These studies use a variety of mathematical tools to extract meaningful patterns from aviation safety databases. These categories will be discussed in the next chapter.

This research sought to first identify different causal factors involved in flight events and then assess their effects on the severity of accidents and incidents. In order to do so, multivariate statistical analyses had to be done over specific databases. In this research Canonical Correlation Analysis (CCA) was applied, which can model a multi-dependent/multi-independent database. By beginning with a descriptive model, a

predictive model can be built from it and used to help identify and prevent potential future casualties. A combination of different flight variables and their effect on outcomes of accidents can be studied using such a model. Before applying CCA, a well-structured database, based on a main database, i.e. NTSB, was needed. This research then improved upon the existing NTSB database by adding more data from other sources to present a complete and comprehensive database prior to running CCA.

This Chapter provided an overview of background research as well as a short explanation of what this research has done. Chapter 2 contains a review of the literature, both in aviation safety and multivariable statistical analysis. Chapter 3 contains the research thesis, data used in that research, and a description of the data preparation process. The application of CCA, its analyses conducted, and the results are described in Chapter 4. Chapter 5 concludes the work.

Chapter 2 – LITERATURE REVIEW

2.1 Risks in Civil Aviation

An air transportation system is complex, containing three main subsystems; airlines, airports, and air traffic control service—all of which have high interaction between each other (Figure 2-1). The aviation system possesses an arrangement of disparate groups conducting activities over broad geographical areas, acting, essentially, without external boundaries (Schmitt & Gollnick, 2016).

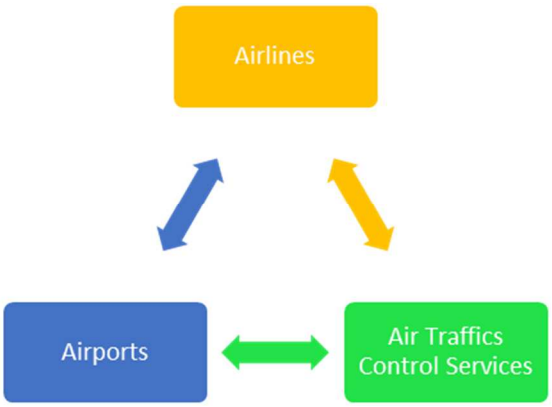

Figure 2-1 Air transport subsystems

Each of these subsystems are complicated technical and technological systems which include human resources, machines, procedures, materials and, of course, management. Errors in each of these subsystems and interactions, which arise from very different components and elements, will produce risks. In order to avoid risk without losing control of the subsystem, it is necessary to accurately control the interaction between

them. Because of such innate intricacy and the serious outcomes of accidents, risks have consistently been one of the highest priorities of the modern air transportation system (Schmitt & Gollnick, 2016).

To eliminate or prevent identified failure in this complex system, the safety program must go beyond the available local information and make causal inferences stretching much further back in time and up into the higher system levels to understand why events happened. There is an expressed need for a model then that can be employed to systematically analyse safety inspection and regulation to display the effects of management decisions. The International Civil Aviation Organization (ICAO), a special organization within the United Nations, has advised its members to create and deploy safety management systems (SMS) to attain sufficient safety precautions for air transportation (LIN, 2011).

As Liou et al. (2007) have indicated, complex factors are involved in airline safety measurements. Those factors include managerial procedures, training, and human resource issues. Therefore the total volume of accidents in a given period of time can be utilized to define a safety index (Liou, Tzeng, & Chang, 2007).

Liou et al. continued the argument that even while numerous studies provide a helpful understanding of aircraft accident/incident data, the fact that accidents are occasional makes it challenging to identify and quickly attend to problems (Liou, Tzeng, & Chang, 2007). They further reference the works of Rose (1992) and Gellman Research Associates (1997) which suggests the rates of civil aviation accidents may not be as effective as a predictive tool for subsequent incidents, and another data set is needed to track the safety levels and tendencies of commercial airlines (Rose, 1992; Gellman

Research Associates, 1997). Beside pilot errors, reasons for accidents that are commonly referenced include the upkeep of aircraft, mistakes in design, and operational deficiencies (McFadden & Towell, 1999). They suggest the implementation of "proactive" safety measures, specifically for tracking accidents due to human error (Liou, Tzeng, & Chang, 2007).

2.2 Risk Measurement in Civil Aviation

The real challenge is how one ought to measure risk. Researchers have pointed out that there are several ways to measure risk (Netjasov & Janic, 2008), based on the kinds that can be recognized in air transportation system (Janic M. , 2000; Netjasov & Janic , 2008):

1) Real risk to an individual: Future circumstances could be a basis for determination of this risk. New technologies in system components are occasionally introduced by incorporation of this risk in decisions.
2) Statistical risk of an event: Mainly insurance companies are concerned about the statistical risk of an event. This is determined based on available statistical data.
3) Predicted risk: Determined by using relevant historical research and is mainly significant for aviation regulators when they are involved in changes in technologies or air traffic patterns.
4) Perceived risk: This is significant for users of the air transportation system and is mainly governed by one's feelings and opinions.

As suggested by Barnett and Wang (2000) "death risk per flight" is the ultimate risk measurement in aviation, and is based on this question: what probability is there that

someone could be killed in a randomly selected non-stop flight? The shortage of this metric is that it neglects that airplane accidents could result in casualties both for on-board and on-the-ground individuals. This risk that threatens on-the-ground individuals is third party risk (Barnett & Wang, 2000; Roelen, 2008). The 9/11 attacks on the World Trade Center towers is a well-known example of such a third party risk (History.com Staff, 2010). There are two famous metrics for this type of risk: "individual risk (IR)" and "group risk (GR)."

- **Individual risk (IR)**

"Individual risk (IR) is the probability – per year – that an imaginary person that is permanently residing at the same location (in the vicinity of the airport) dies as a direct result of an aircraft accident" (Roelen, 2008).

- **Group risk (GR)**

"Group risk (GR) is defined as the probability - per year- that a group of more than a particular number of persons (denoted with *N*), other than occupants of the aircraft, dies as a result of a single aircraft accident" (Roelen, 2008).

Roelen (2008) continues the discussion on how this risk is illustrated. GR is regularly spoken to in a chart plotting the likelihood of an accident occurring with *N* or more fatalities. In the social sciences, GR is often labelled disaster probability as it addresses both the probability as well as the size of the event. (Roelen, 2008).

When quantifying aviation risk using statistical data of accidents, rarity should be considered. If a dice is tossed five times only and all of the outcomes were 6 for example, this cannot be evidence of having unfair dice (Czerwinski & Barnett, 2004). The quantity of fatal flights has been so low that it is harmful to utilize as the main marker of risk in air

transportation. To be more comprehensive, accidents and incidents without fatalities ought to be incorporated into the contemplations on the development and construction of a model of causation. (Roelen, 2008). That is what has been implemented in this research, which will be presented in detail in subsequent chapters.

2.3 Accident Causation Theories

Roelen (2008) discusses that risk from air transport is only identified with accidents. In order to increase safety by reducing the probability of accident occurrences, researchers have attempted to build a framework that illustrates the causes of accidents. Some of these speculations have been exceptionally powerful. In transit, we consider accidents how they are represented through numerical models. Hence this segment gives a concise review of some of the most critical accident causation speculations and finishes up with their implications for risk model necessities (Roelen, 2008).

This section will discuss relevant safety theories. The safety models include the SHELL Model of human factors, the Swiss Cheese Model, and the Domino Effect Model.

2.3.1 The SHELL Model of Human Factors in Aviation Safety

The SHELL Model is characterized as "the relationship of human factors and the aviation environment" (Reinhart, 1996). Edward (1972) began a new concept named "SHELL Model," whereby the acronym stems from its components (Software, Hardware, Environment, and Live-ware) (Perezgonzalez, Lee , LI , & ZiZhang, 2011). In 1975, Hawkins added an *L* by introducing a second Live-ware into the initial model, suggesting the title "SHELL Model" (Hawkins F. , 1987).

Edwards's SHEL Model (1972) and Hawkins's SHELL Model (1975) are differentiated mainly through Hawkins emphasis on the significance of another "Liveware" that is human, and proposed a diagram to better illustrate the interactions between different systems (Hawkins F. , 1987).

It was widely noted that the vast majority of the accidents are caused by human blunders; however, failures in the aircraft mechanical systems in aircraft support has brought on the abatement of various innovative gear developments (Hawkins F. , 1987). Besides, in view of the human component, each person, either who operates or supports some portion of aviation, has singular abilities and confinements. Therefore, many nations have considered interaction of each SHELL component as a basis for their training programs to secure safety (Hawkins F. , 1987; Perezgonzalez, Lee , LI , & ZiZhang, 2011) (Figure 2-2 SHEL Model).

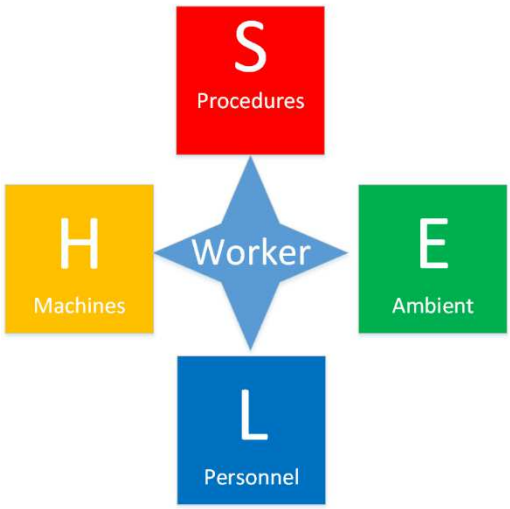

Figure 2-2 SHEL Model

The fundamental components in the model can be distinguished as hardware—which

involves different non-human, mechanical elements, including the facilities, aircraft equipment, and the other physical assets of aviation; and software—all procedures implemented, other resources such as organizational hierarchies and policy, rules, instruction manuals, and notices. Another component is the environment—which involves exterior elements that impact the individuals who are working including weather, noise, and other potential disturbances, as well as larger systemic factors like the socio-political climate and economy. The live-ware incorporates interpersonal components like teamwork, communication, leadership, and norms. The central live-ware is contained within the human actors and includes things such as workers' KSAs, culture, and relative stress levels. This live-ware sits at the core of the SHELL Model and the different segments that stem from it (Hawkins F. , 1987; Perezgonzalez, Lee , LI , & ZiZhang, 2011).

2.3.2 Swiss Cheese Model on Aviation Safety

As indicated by Wiegmann & Shappell (2003), the Swiss Cheese Model—a simple and effective model created by Reason in 1990—was widely accepted across the industry as the authoritative accident model. It provided a way to see how aviation can operate effectively or drift into failure (Wiegmann & Shappell, 2003). According to the Swiss Cheese Model (also called the Reason model) accidents occur because of the confluence of a number of factors, which are not enough on their own to cause a systemic break. Reason contends that complex systems like the aviation system are layered with defences, and that in-depth, single-point failures seldom lead to consequences. Failures of equipment or mistakes in operations are often just small triggers and the result of larger systemic issues. Ruptures across the defences are a postponed result of choices made

system-wide, which stay lethargic until their potential is met by a concurrence of events (Wiegmann & Shappell, 2003).

As mentioned by ICAO (2012) in the Safety Management Manual (SMM), during these situations human error or active operational mistakes set off a breach of the layered safety defences, but the conditions surrounding the failure have to be conducive to the triggering event. As the idea presented by Reason attests, all accidents include active and latent conditions in concert (ICAO, 2012).

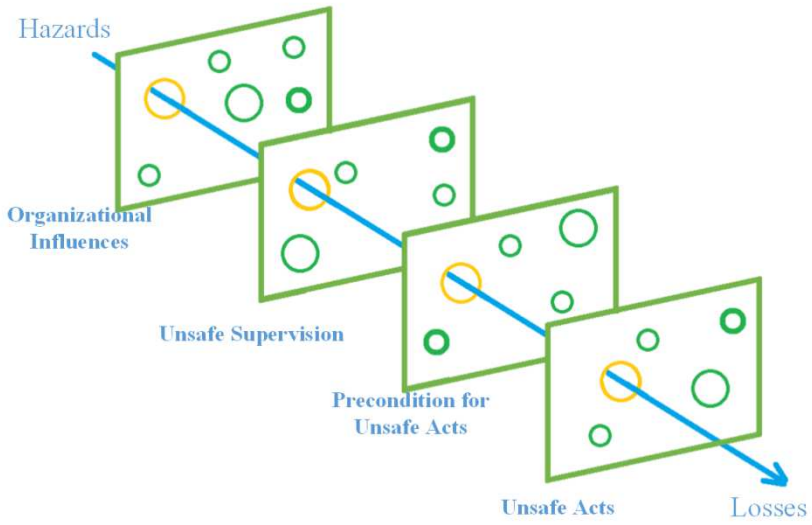

Figure 2-3 Swiss Cheese Model, (Modified from (Reason, 1990))

The Swiss Cheese Model (Figure 2-3) of accident causation presents the system of human defences as a circumstantially well-aligned series of randomly-holed Swiss cheese slices organized vertically and parallel with crevices in the middle of every cut. Four levels of failure could be considered as the source of accidents if they could be traced

back, as hypothesized by Reason. The four groups are directly quoted as follows (Wu, et al., 2009):

- "Organisational influences,
- Unsafe supervision,
- Preconditions for unsafe acts, and
- The unsafe acts themselves"

In this model, an organization's attempts at safety are displayed as a progression of obstructions, represented as the slices of cheese. The holes in each slice represent individual weaknesses in parts of the system. All slices contain differently sized and placed holes. Failures occur when the holes align—allowing what seems to be an impenetrable wall of safety to penetrated by a hazard, producing an accident. (Lundberg, Rollenhagen, & Hollnagel, 2009). The system only allows accidents when the alignment is just right, a confluence of discrete events that entail a tragic event.

2.3.3 The Domino Effect Theory on Aviation Safety

The Domino Safety Theory was created by H. W. Heinrich and Alfred Lateiner (1931), innovative leaders in the field of industrial safety. It provides a realistic representation of the occurrence of industrial injuries as well as how they can be prevented. The Domino Effect Theory is also called as the Cause-Effect Theory or Causal Model.

This theory supposes that an accident occurs as the end result of a sequence of events; it is a chain response. To better understand the model, imagine five dominoes consecutively, the main domino is foundation—a worker's way of life and identity. The second domino is close-to-home qualities—a worker's state of mind, level of learning,

and physical and mental conditions (Figure 2-4).

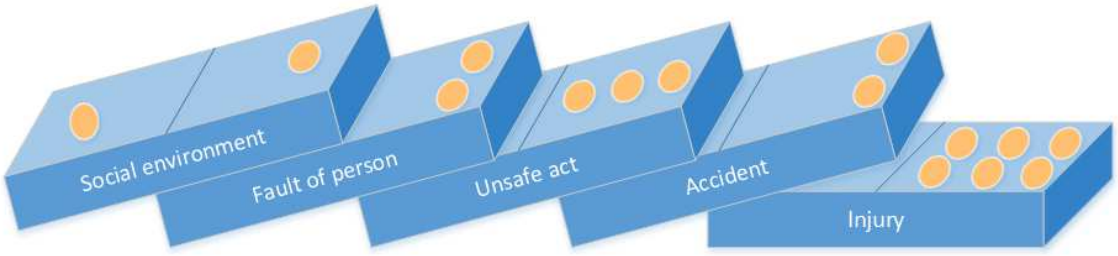

Figure 2-4 Domino Model of Accident Causation (Modified from Heinrich, 1931)

The third domino is the unsafe action taken by the worker. The fourth domino is the accident, precluded by the unsafe action. The fifth domino is the damage—the outcome or injury caused by the series of events. The Domino theorists argue that for any given incident, very little can be done about a worker's foundation and individual qualities. The unsafe acts and unsafe conditions are the two dominos that must be targeted. The worker should be stopped immediately once an unsafe act is detected, at which point a series of actions should be undertaken. First the situation must be examined, with the hopes of finding a safer means to accomplish the task. Then the worker must be trained to perform the task accordingly. Intermittent checks of his or her performance should be conducted, with retraining as needed, and finally if nothing else has brought about corrected actions, then the worker must be disciplined. If unsafe conditions are found, it needs to be corrected or workers need to be informed of it. Heinrich insists that foremost the employer should be assumed responsible. He states that managers concerned about safety will take the measures necessary to ensure his or her employees do as instructed to remove the unsafe conditions (Forest Resources Association, 2002). Heinrich recommends firm supervision as a solution for such non-compliance, including

remediation and discipline. Many perspectives on human error and failure exist. Each of them is formed assuming particular underlying causes for failure. It would be fair to say that there are nearly as many models on failure as there are researchers addressing it (Senders & Moray, 1991; Wiegmann & Shappell, 2001).

2.4 Aviation Safety Studies

Air transportation is one of the fastest growing modes of transportation, 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.

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 W.-K. , 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 civil aviation risk and safety has been the focus of many studies, and possible approaches for improving the safety of general aviation have been put forth (Janic M. , 2000; Li & Baker, 2007). In

general, these researches can be divided into three main groups, as shown in Figure 2-5 (Kashef, 2016).

Figure 2-5 Summary of Research on Civil Aviation Safety

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."

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 hijacking.

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 recognize and alleviate aviation errors caused by people (Nelson, 1997). 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 connections between team situation awareness (SA) and the sharing of information, and propose a method to reduce human error. The authors also compare pilot SA and air traffic controller (ATC) requirements (Jin, Sun, & Kong, 2010). 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, namely 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 cost reduction, 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 analyse 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 connection between delayed flight and risk, and proposed 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 affecting 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, avoidable or unavoidable, and further differentiates causal factors based on accident type—i.e., whether they can be attributed to human or mechanical failures, weather hazards, sabotage, or military operations (Jiao, Zhao, & Wang, 2009).

Spouge (2004) further discusses the benefits of causal analysis and argues that safety managers and policy makers must understand the causes of accidents and assess 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, incorporating multiple safety performance indicators, in order 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 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 scarcity of events. In addition, 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 a 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, the flight phase, and aircraft type 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 may seem 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. To do so, firstly, a reliable database is needed and secondly, an analysis tool is required to analyse the data. In statistics there are different such tools that help researchers do analysis. Whether dealing with single or multi-variable will change appropriate statistical tools. In our research we will be dealing with multi-factors as mentioned, therefore, investigating the relations between these factors—and specifically as dependent and independent variables—is the main focus of this research. Thus in the next section, a thorough review of the literature on multivariable statistical analysis, its history, tools, and mathematical fundamentals will be reviewed and discussed. Afterward its application on the database prepared in this research will be presented.

2.5 Statistical Methods

2.5.1 Introduction

The most straightforward approach to deal with statistical modeling is to relapse one variable upon another. The simplest way to make such models is via regression. First defined by Sir Francis Galton (1885), one variable regression (simple linear regression) is a model with one independent and one dependent variable. However, in the real world, each phenomenon is related to more than one variable—the more variables the greater influence over the response. Multiple regression was created to answer these situations. Multiple regression is an extension of simple linear regression. Multiple regression allows one to fit a model to data when a single variable (y) is believed to depend on a number of other variables. It is used when we want to predict a variable's value based on the value of two or more other variables (Borga, 2001).

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$ = amounts of response variable

$$\begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1q} \\ x_{21} & x_{22} & . & . & . & x_{2q} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & . & x_{nq} \end{bmatrix}$$ = amounts of independent variables $(x_1, x_2, \dots, x_q)$.

Multiple regression tries to relate a vector of independent data to a solitary response variable based on following equations; the goal is to find regression coefficients (Salkind, 2010).

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \epsilon_2$$

.

.

.

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_q x_{nq} + \epsilon_n$$

The number of *x*'s is denoted by q while βqs are regression coefficients. This linear model requires no extra terms to predict y; the remaining variation in *y* is arbitrary and unpredictable. If the mean of εi is zero ((εi)=0) then,

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}$$

In multiple regression, the slope of the linear relationship between the dependent variable and the part of a predictor variable is the regression coefficient. This slope is an indicator of covariance between independent and dependent variable $Cov(x_i, y)$ (Rencher, Methods of Multivariate Analysis, 2002).

Variances and co-variances are all paired variables, shown as the following matrix:

$$S = \begin{pmatrix} S_{yy} & S_{y1} & S_{y2} & . & . & . & S_{yq} \\ S_{1y} & S_{11} & S_{12} & . & . & . & S_{1q} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ S_{qy} & S_{q1} & S_{q2} & . & . & . & S_{qq} \end{pmatrix} = \begin{pmatrix} S_{yy} & S'_{yx} \\ S_{yx} & S_{xx} \end{pmatrix}$$

Where Syy is the variance of y, Syj is the covariance of y and Sjj is the variance of xj, Sjk is the covariance of xj and xk and $S'_{yx} = (S_{y1}, S_{y2}, \dots, S_{yq})$.

One of the most important factors in regression is correlation. Correlation is a measure used to determine the relationship between two variables (Nelson, 1997). To be precise, correlation is a statistical tool that frequently refers to a linear relationship between two variables. The next section will be more specific about correlation details.

2.5.2 Correlation

Mukaka (2012) mentions inaccurate use of the term correlation in verbal communication. However in the academic environment, the term correlation is used as quoted: “an association, connection, or any form of relationship, link or correspondence”. This broad definition itself caused some misuse of the statistical term “correlation” among researchers. This misuse is to such an extent that some statisticians have wished the method had never been concocted (Mukaka, 2012). Correlation is one of the most seasoned ideas in statistics and can seemingly be followed back to Gauss in 1823 (Rodgers & Nicewander, 1988). Although the fundamental thought of measuring the relationships between variables was expected generously before 1885, Sir Francis Galton was the first one who completed the theory of bivariate by coining the term "regression" (Rodgers & Nicewander, 1988). A decade later, Karl Pearson formalized the method and developed the index that we still use to measure correlation, Pearson's *r* (Pearson, 1920). Pearson's *r* was the principle formal correlation measure; it is still the most broadly utilized measure of relationship today. Rodgers & Nicewander (1988) have summarized milestones in the history of correlation and regression as directly quoted and shown in Table 2-1.

Table 2-1 Points of interest in the historical backdrop of Correlation and Regression (Rodgers & Nicewander, 1988)

Date	*Person*	*Event*
1823	Carl Friedrich Gauss, German mathematician	Developed the normal surface of *N* correlated variates.
1843	John Stuart Mill, British philosopher	Proposed four canons of induction, including concomitant variation.
1846	Auguste Bravais, French naval officer and astronomer	Referred to "une correlation," worked on bivariate normal distribution.
1868	Charles Darwin, Galton's cousin, British natural philosopher	'All parts of the organization are . . . connected or correlated."
1877	Sir Francis Galton, British, the first biometrician	First discussed 'reversion, the predecessor of regression.
1888	Sir Francis Galton	First referred to “regression."
1895	Karl Pearson, British statistician	Defined the (Galton-) Pearson product-moment correlation coefficient.
1920	Karl Pearson	Wrote “Notes on the History of Correlation."
1985		Centennial of regression and correlation

Any measurable relationship between two arbitrary variables could be called a "Correlation". Correlation refers to any of an expansive class of statistical relationships including reliance; however, it often refers to the degree to which two variables have a linear relationship with each other. As mentioned, correlations are valuable because they can demonstrate a prescient relationship that can be abused practically speaking.

The Pearson product-moment correlation coefficient, which is invariant to linear transformations of either variable, is a dimensionless index (Rodgers & Nicewander, 1988). To calculate this index we need to divide the covariance of the two variables by the result of their standard deviations.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{[\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2]^{\frac{1}{2}}}$$

Mukaka (2012) discussed the possible values for correlation and what those values might mean. A correlation coefficient is between -1 to +1. Zero shows that no linear relationship exists between two continuous variables, while a correlation coefficient of −1 or +1 demonstrates an impeccable linear relationship. The calculated values of correlation, which could be −1 and +1, show the strength of relationship. As we move closer to ±1 in the spectrum, the correlation coefficient will be stronger. A positive coefficient means a direct relationship between variables (i.e., increasing the value of one variable, causes the increase in value of the other one) (Mukaka, 2012). However the

negative coefficient means an inverse relationship between variables (Mahdavi Damghani, 2013).

In spite of correlation simplicity, the technique has a few confinements, and the coefficient is frequently erroneously interpreted. The Pearson coefficient is precise only if the correlation relationship is both linear and simultaneous; subsequently it can't catch if an adjustment of X in period 1 is correlated with an adjustment of Y in period 2. Also, it has been represented that the correlation result can be essentially influenced by anomalies in the data, which might speak to off base estimations or great occasions (Anscombe, 1973). Anscombe (1973) likewise distinguishes four information sets with indistinguishable relationship and other factual properties, (for example, skew and mean) which can have totally diverse disseminations of data. This exhibits that there is a point of confinement to how useful relationship can be. Along these lines, correlation is plainly an exceptionally helpful statistical method; however, it is constrained by presumptions fundamental to the estimation, and thus ought to be connected when those presumptions hold (Mahdavi Damghani, 2013).

2.5.3 Multivariable Statistical Analysis

To evaluate the effect of multiple independent variables on the correlation between two sets of variables, a multivariate statistical analysis tool is necessary. Usually researchers in different fields collect measurements on several variables. Multivariate analysis (MVA) is a set of advanced statistical, computational, and pattern recognition techniques (i.e., multivariate regression, multivariate analysis of variance, machine learning, etc.) that can be used when several measurements are involved in the database simultaneously. MVA technologies provide a theoretical improvement over univariate

techniques, which examine each acquired measurement individually. Unsurprisingly, the heft of the uses of multivariate procedures have been in the behavioral and biological sciences, complex transportation systems analysis, and so on. (Rencher, 2002) .

2.5.3.1 Advantages of Multivariate Statistics

Researchers have discussed advantages of multivariate statistics over single variable regression analysis. McGarigal et al. (2000) described some of those advantages as follows:

- Especially for the natural systems, it reflects more precisely their true multidimensional, multivariate nature.
- While working with large data sets with too many variables, it provides a way to handle the data sets.
- It gives a way to summarize redundancies in the data set.
- An "optimal" combination of variables can be created by applying the rules provided by this method.
- It has a solution to perform multiple comparison problems by controlling error rate from experimentation.
- “Provide[s] a means of detecting and quantifying truly multivariate patterns that arise out of the correlational structure of the variable set”.
- “Provide[s] a means of exploring complex data sets for patterns and relationships from which hypotheses can be generated and subsequently tested experimentally”. (McGarigal, Cushman, & Stafford, 2000).

2.5.3.2 Classification of Multivariate Techniques

There are four basic types of (continuous) multivariate data, and for each there are possible analysis techniques proposed in the literature. One may consider this an over-simplification, preferring elaborate tree diagrams of data structure (Rencher, 2002). However, most of data sets can fit into one of these classes, and the simplicity of this structure makes it easier to remember. Table 2-2 shows each data structure with models, techniques and objectives and Figure 2-6 show data structures, schematically.

Table 2-2 Multivariate techniques classification based on data structure, modified from (McGarigal, Cushman, & Stafford, 2000)

Data Structure	Model	Techniques	Objective
One set, more than two variables	y1 + y2 + ... + yi	(Unconstrained Ordination) PCA, MDS, CA, DCA, NMDS	Extract gradients of maximum variation
One set, more than two variables	y1 + y2 + ... + yi	Cluster Analysis (Family of techniques)	Establish groups of similar entities
Two sets; one grouping variable,	y1 + y2 + ... + yi = x	Discrimination (MANOVA,	Test for & describe differences among

more than two discriminating variable		MRPP, ANOSIM, Mantel, DA, LR, CART, ISA)	groups of entities or predict group membership
Two sets; more than 2 response variables, more than two explanatory variables	y1+ y2+ ... +yi = x1+ x2+ ... + xj	Constrained Ordination (RDA, CCA, CAP)	Extract gradients of variation in dependent variables explainable by independent variables

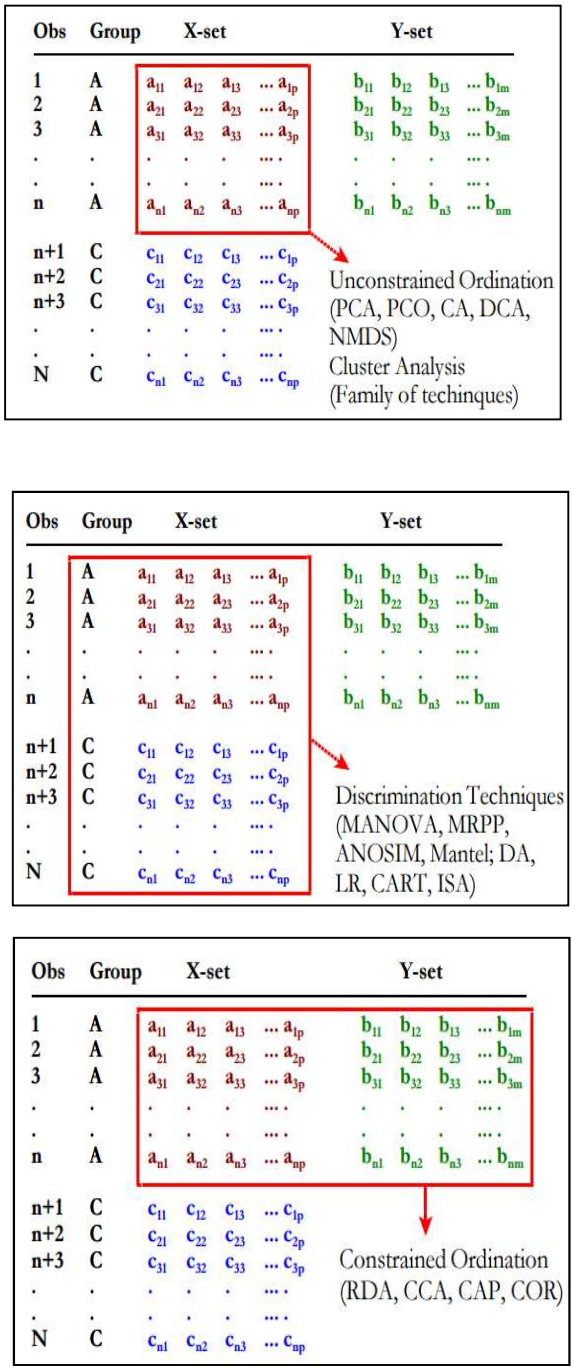

Figure 2-6 Data Structure for Different Multivariate Techniques, modified from (McGarigal, Cushman, & Stafford, 2000)

2.5.4 Canonical Analysis

Canonical analysis can be considered a part of the family of regression approaches. As discussed by Kuehl (1999), when dealing with predictor (independent) variables and dependent variables, regression analysis can be used as a quantifier of the relationship between these two groups of variables. The analysis also calculates the coefficient of correlation *r*, coefficient of determination r^2, and the standard regression coefficient β (Kuehl, 1999). Although, multiple regression is the most widely used method for finding relationships between several quantities, it is restricted to a single dependent variable. In addition to multiple regression, canonical analysis is a generalizing of all those approaches. It finds the ideal linear combination of predictor data vectors that will clarify the most changes in predicted data vectors. A relationship between two sets of variables (one as predictor or independent and the other one as predicted or dependent) can be captured using canonical variate analysis. The canonical correlations ρ_1, ρ_2,… and the sets of canonical weights will then be calculated as well.

2.5.4.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is one of the most common methods of multivariate analysis and multi-dimensional aims to determine the relationship between variables. CCA, first developed by H. Hotelling (Hardoon, Szedmak, & Shawe-Taylor, 2004), is a surprisingly little used technique.

CCA is a multivariate technique that measures the linear combination (i.e. a weighted sum) between two multi-dimensional variables and aims to quantify and identify the associations between them. This linear combination of the dependent variables and a

linear combination of the independent variables maximizes the correlation between the two.

In each research unit we regularly measure two types of variables against each other, i.e. variables related to poverty such as income, expenses, and access to water relate to variables like "distance to water point, time taken to water point" (Mkondiw, Jumbe, & Wiyo, 2013). These two sets of variable can be shown with two vectors:

$$Y' = (y_1, y_2, \dots, y_p) \qquad X' = (x_1, x_2, \dots x_q)$$

The main problem in CCA is defined as finding two sets of basis vectors, for *x* and *y* respectively, such that "the correlations between the projections of the variables onto these basis vectors are commonly maximized" (Borga, 2001). The initial step is to find two linear combinations with the largest correlation:

$$U_k = a_1 x_1 + a_2 x_2 + \dots + a_q x_q$$

$$V_k = b_1 y_1 + b_2 y_2 + \dots + b_p y_p$$

In the next step the pair of linear combinations with the largest correlation among all pairs are determined with the condition that they are not correlated with the pair selected initially. The pairs of U_k, V_k are called canonical variables and correlation between them is dubbed canonical correlations, which measures the amount of association among variable sets X' and Y (Johnson & Wichern, 2002). "Canonical correlation is an extension of multiple correlation, which is the correlation between one *y* and several *x*'s and is often a useful complement to a multivariate regression analysis" as quoted from Fan & Konold (2010).

2.5.4.2 CCA Terminology

Sherry & Henson (2005) outlined different aspects of CCA, its application, and important terms used in this method. For interpreting and understanding the CCA, it is important to understand and learn its language, and, along these lines, compose a succinct results area for manuscript. A large portion of the insights within a CCA have univariate analogs, thus it would be useful if comparative insights would have comparative names crosswise over investigations. Unfortunately, this is regularly not the case. At the risk of building up a few shared characteristics with different analysis, i.e. multiple regression, we present the following implications of pertinent CCA statistics (Sherry & Henson, 2005):

- **Canonical correlation coefficient:**

After making two manufactured variables (canonical variables, $U_i \; and \; V_i$), Pearson r, which is called here *canonical correlation coefficient*, is calculated. This coefficient is between 0 and 1 and can't be negative.

Canonical correlation coefficient magnitude express the degree of linear correlation between $U_i \; and \; V_i$. These dimensionless coefficients remain constant under nonsingular linear transformations of the variables of either/or both set.

- **Squared canonical correlation:**

Since the canonical variables speak to the watched independent and dependent variables, squared canonical correlation demonstrates the measure of shared variance between the variable sets.

- **Canonical function**

A set of standardized canonical coefficients that are in linear equations (Ui and Vi) are called canonical function.

$$U_i = a_1x_1 + a_2x_2 + \ldots + a_qx_q$$

$$V_i = b_1y_1 + b_2y_2 + \ldots + b_py_p$$

- **Standardized canonical function coefficients**

These coefficients are used in the linear equations discussed in the previous section. "These weights are applied to the observed scores in Z-score form" (Sherry & Henson, 2005).

- **Structure coefficient**

This parameter is basically the Pearson's *r* between the canonical function scores and an observed variable for the variable's set. Pearson's *r* statistics are simple structure coefficients that range from -1 to +1 including both negative and positive 1. They help in understanding the way manufactured variables are structured and can illuminate the interpretation of such variables. The structure coefficients calculated in a factor analysis structure matrix can be compared to these coefficients (Henson R. K., 1999; Sherry & Henson, 2005; Courville & Thompson, 2001).

2.5.4.3 Advantages of CCA

CCA as a multivariate technique has different advantages that can be listed as follows (Sherry & Henson, 2005):

1) In general for multivariate techniques including CCA, the probability of committing Type I error is limited during research (Thompson, 1991). In any research, type I error is occasionally called "experimentwise error" (Sherry &

Henson, 2005). When a lot of statistical tests are carried out on the same variables in a data set, each test has type I error (for example alpha=0.5). Multivariate techniques do not need to conduct these tests or compare variables simultaneously that minimize type I error.

2) Multivariate techniques, for example CCA, might best respect the reality of research. Such techniques most scrutinize regularly researched variables that potentially have numerous causes and impacts. Findings based on research that independently considers particular causes and effects may distort the complex reality of research in different situations.

3) CCA technique can be utilized rather than other parametric tests in diverse occasions, making it an imperative method to learn as well as a thorough technique. As many researchers have mentioned (Henson R. K., 1999; Knapp, 1978), in special cases CCA can be used instead of other statistical techniques.

2.5.4.4 Appropriate Utilization of CCA

The best usage of CCA is when the goal is to examine two sets of variables and evaluate the relationship between them. Before starting the analysis and putting variables into two groups, there should be some justification for why those variables are treated together in a set (Sherry & Henson, 2005). As an example, there might be 10 variables relating to flight and others related to the severity of an incident once it has occurred (i.e. number of injuries, aircraft damage, etc.). The research question then would be whether or not there is a relationship between flight variables and event variables (Kashef, 2016). On the other hand, if the problem is such that only one criterion is involved, then multiple regression would be more applicable. In cases like our research, where multiple variables

exist in both sets, CCA is the most suitable analytical methodology (Sherry & Henson, 2005).

The basis for which set works as predictor is the researcher's judgment and understanding both about the predictive causality as well as which set would have the predicted role. Having said that, CCA and its nature as a correlational tool ultimately makes such a declaration arbitrary and rather recommends the researcher take care when inferring about causal relationships outside of an experimental design (Sherry & Henson, 2005). Since CCA looks at the relationship between weighted canonical variables (canonical variables for *x*'s and canonical variables for *y*'s), taking into account the connections between variables inside the sets, CCA can be conceptualized as a basic bivariate correlation (Pearson r) between the two synthetic variables (canonical variables).

As a general illustration Figure 2-7 shows an assumed CCA model with *q* indicator (independent) variables and *p* dependent variables. A combination of each set's variables should be generated—this is called a canonical variate or synthetic variable. In CCA, a linear equation is applied to create those variables. The equation is between independent

variables that have synthetic variable for IVs (Xs); the same applies to canonical variates for Ys.

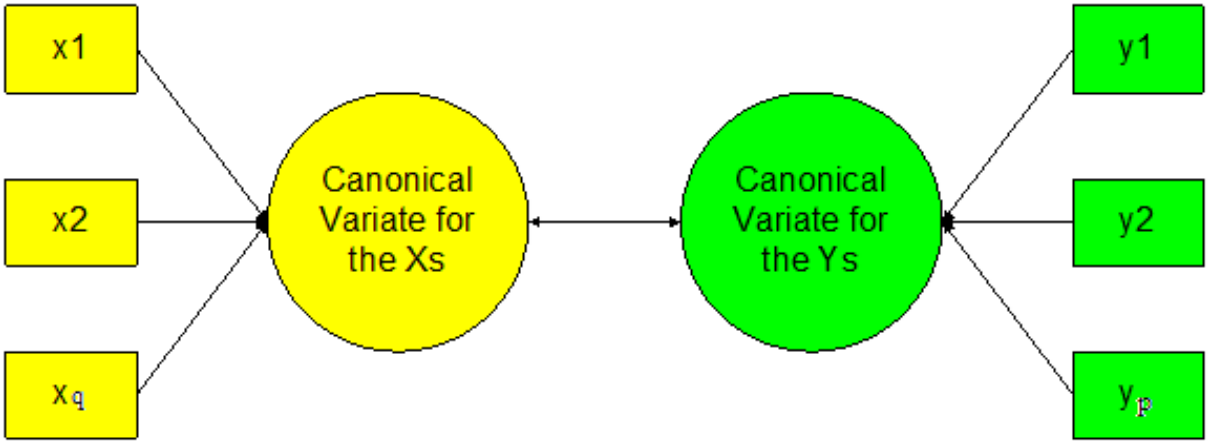

Figure 2-7 Conceptual model for canonical correlation, modified from (Sherry & Henson, 2005)

This utilization of linear equations is specifically and practically equivalent to the utilization of linear equations in the more well-known multiple regression, in which beta (β) weights are duplicated with watched scores (in Z score form) then summed to yield engineered anticipated scores (i.e., $Y' = \beta_1 x_1 + \beta_2 x_1$). Using institutionalized weights closely resembling beta weights, CCA makes two direct comparisons: one for the indicator variables and one for the predicted variables. The results of these comparisons are the two manufactured variables, as in Figure 2-7.

2.5.4.5 CCA Assumptions

As with all statistic tests, suitable utilization of CCA assumes some fundamental points presented below (Tabachnick & Fidell, 1996):

Normality: As in the case with multivariate regression, MANOVA, and so on, CCA requires the multivariate normal and a homogeneity of variance assumption for valid inference.

Linear relationship: CCA assumes a linear relationship between the canonical variates and each set of variables. The event of a curvilinear relationship will reduce the adequacy of the analysis. The main goal of this analysis method is to find "linear relationships" between variables within a set and between canonical variate pairs (French, Chess, & Santos, 2005). Hence linearity is imperative in a practical sense because Pearson's r, which is basic to most parametric statistical procedures (Graham, 2008), just catches the linear relationship among variables. Nonlinear components of these relationships are not recognized and so are not 'captured' in the analysis. Pearson's *r* does not fully account for the actual non-linear relationship between two variables (Warner, 2008).

Outliers: As with all least squares procedures, outliers can significantly influence the extents of correlation coefficients. Since CCA is based on correlation coefficients, they can likewise seriously affect canonical correlations. In order to minimize the impact of one or two outliers, it is important to use a larger sample size (Hardoon, Szedmak, & Shawe-Taylor, 2004).

2.5.5 Theoretical Foundations

2.5.5.1 Multiple Correlation Review

As mentioned before, since CCA is an extension of multiple correlation, to build our theoretical foundation we will review multiple correlation as the base of CCA. The

matrices can be used to show the sample co-variances and correlations among *y* and x_i $(i = 1\ to\ q)$ (Rencher & Christensen, 2012):

$$S = \begin{bmatrix} S_y^2 & S'_{yx} \\ S_{yx} & S_{xx} \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & r'_{yx} \\ r_{yx} & R_{xx} \end{bmatrix}$$

Which in them,

$$S' = (s_{y1}, s_{y2}, \dots s_{yp})$$

S_{xx} = Matrix of sample covariance of the *x*'s.

$$r'_{yx} = (r_{y1}, r_{y2}, \dots , r_{yp})$$

R_{xx}= Matrix of sample correlation of *x*'s.

To express R^2 in terms of sample variances, co-variances, and correlations, the following equation is used:

$$R^2 = \frac{S'_{yx}.\ S_{xx}^{-1}.S_{yx}}{S_{yy}} = r'_{yx}.R_{xx}^{-1}.r_{yx}$$

In multiple correlation, "R can be defined as the maximum correlation between *y* and a linear combination of the *x*'s" (Rencher & Christensen, 2012).

$$y \longrightarrow x_1 \quad x_2 \quad x_3 \quad \dots x_q$$

2.5.5.2 Mathematics of CCA

It is the goal of CCA to measure the relationship between two sets of variables (several *y*'s and several *x*'s). A sample covariance matrix for mixed vector (Holmström & Clemesha, 2008):

$$(y_i, i = 1\ to\ p\ , x_j, j = 1\ to\ q)$$

is:

$$S = \begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix}$$

For the random vectors $Y' = (y_1, y_2, \dots, y_p)$ and $X' = (x_1, x_2, \dots x_q)$, let

$$E\left((x_1, x_2, \dots x_q)\right) = \mu_x \text{ , } Cov\left((x_1, x_2, \dots x_q)\right) = S_{xx}$$

$$E\left((y_1, y_2, \dots, y_p)\right) = \mu_y \text{ , } Cov\left((y_1, y_2, \dots, y_p)\right) = S_{yy}$$

$$Cov\left((y_1, y_2, \dots, y_p), (x_1, x_2, \dots x_q)\right) = S_{xy} = S_{yx}$$

Then, S can be calculated as follows:

$$\text{S} = \begin{bmatrix} E(X' - \mu_x)(X' - \mu_x)' & E(X' - \mu_x)(Y' - \mu_y)' \\ E(Y' - \mu_y)(X' - \mu_x)' & E(Y' - \mu_y)(Y' - \mu_y)' \end{bmatrix} = \begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix}$$

Where S_{yy} is the $p \times p$ sample covariance matrix of the y

s, $\boldsymbol{S}_{yx}$ is the $p \times q$ matrix of sample co-variances between the y's and the x's, and $\boldsymbol{S}_{xx}$ is the $q \times q$ sample covariance matrix of the x's (Rencher & Christensen, 2012; Weenink, 2003).

If U and V are defined as two linear combinations of y's and x's ($U_i = a_i' y \; and \; V_i = b_i' x$) and if $s = \min(p, q)$ then there are s canonical correlations $r_1, r_2, \dots, r_s$ corresponding to the s pairs of canonical variates U_i and V_i ($r_i = r_{V_i U_i}$) as shown in Table 2-3.

Table 2-3 Pairs of Canonical Variates

r	Linear combination for U	Linear combination for V
$\boldsymbol{r_1}$	$U_1 = a_1'y$	$V_1 = b_1'x$
$\boldsymbol{r_2}$	$U_2 = a_2'y$	$V_2 = b_2'x$
…	…	…
$\boldsymbol{r_s}$	$U_s = a_s'y$	$V_s = b_s'x$

Based on above definitions, $Var(U), Var(V)\ and\ Cov(U,V)$ can be measured as follows:

$$Var(U) = a'Cov(X')a = a'S_{xx}a$$

$$Var(V) = b'Cov(Y')b = b'S_{yy}b$$

$$Cov(U,V) = a'Cov(X',Y')b = a'S_{xy}b$$

The problem is to find maximum correlation between U and V based on the coefficient vectors of a and b. With the following equation:

$$Corr(U,V) = \frac{a'S_{xy}b}{\sqrt{(a'S_{xx}a)(b'S_{yy}b)}}$$

The coefficient vectors $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ in the canonical variates $U_i = \boldsymbol{a_i'y}$ and $V_i = \boldsymbol{b_i'x}$ are the eigenvectors of these two matrices (Rencher, 2002):

$$(S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} - r^2I)a = 0$$

$$\left(S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx} - r^2I\right)b = 0$$

Since the matrix $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$ is not symmetric and algorithms for eigenvalues and eigenvectors often accept symmetric matrices based on symmetry in S_{yy}^{-1}, $S_{yx}S_{xx}^{-1}S_{xy}$, and

$U'U = S_{yy}$ equations, the following relationship can be used to produce a symmetric matrix:

$$(U')^{-1}S_{yx}S_{xx}^{-1}S_{xy}U^{-1}$$

This symmetric matrix has the same eigenvalue as $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$ but has eigenvectors Ua_i.

As a result, the pq co-variances between the *y*'s and *x*'s in S_{yx} have been supplanted by s = min (p, q) canonical correlations. These condense the relationship in the middle of *y* and x. The smallest eigenvalues can be slighted to accomplish considerably more disentanglement. The significance of every eigenvalue can be judged by its relative size (Rencher, 2002):

$$\frac{r_i^2}{\sum_{j=1}^{s} r_j^2}$$

The canonical correlations can likewise be obtained from the divided correlation matrix of the *y*'s and *x*'s,

$$R = \begin{bmatrix} R_{yy} & R_{yx} \\ R_{xy} & R_{xx} \end{bmatrix}$$

$R_{yy}, R_{yx}, R_{xy}\ and\ R_{xx}$ denote $p \times p$ sample matrix of the *y*'s, $p \times q$ sample matrix correlation between *y*'s and *x*'s, and $q \times q$ sample correlation matrix of *x*'s, respectively. The following equations then produce *c* and *d* eigenvectors:

$$(R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy} - r^2 I)c = 0$$

$$\left(R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx} - r^2 I\right)d = 0$$

There are relationships between *c* and *d* with a and b:

$$c = D_y a \; and \; d = D_x b$$

D_y and D_x denote $diag\left(S_{y1}, S_{y2}, \ldots, S_{yp}\right)$ and $diag\left(S_{x1}, S_{x2}, \ldots, S_{xp}\right)$.

In fact, *c* and *d* are standardized coefficient vectors and can be applied to standard variables.

$$u = a'(y - \bar{y}) = a' D_y D_y^{-1}(y - \bar{y}) = c' D_y^{-1}(y - \bar{y})$$

$$= c_1 \frac{y_1 - \overline{y_1}}{s_{y1}} + c_2 \frac{y_2 - \overline{y_2}}{S_{y2}} + \cdots + c_p \frac{y_p - \overline{y_p}}{S_{yp}}$$

2.5.5.3 Canonical Correlations Properties

Rencher (2002) mentions two attributes of canonical correlations:

1. Canonical correlations will not be affected by changes of scale. For instance, canonical correlations are the same whether Celsius or Fahrenheit measures temperature in a given problem, but the corresponding eigenvectors will not remain the same. Both simple and multiple correlations share this trait.
2. Correlations will be calculated in order which means r_1 will be the strongest correlation between the linear combination of *y* and x. As a result, the absolute value of the simple correlation between any *y* and any *x* will be exceeded by r_1. It also exceeds all of the multiple correlations between any *y* and *x*'s or the other way around (Rencher, 2002).

2.5.5.4 Canonical Correlation Applications

The objective of CCA can include any one or more of the following (Lee, Katz, & Restori , 2010):

- Determining if two sets of variables for each object (individual) are linearly correlated.
- Determining which variables in each of the two sets contribute the most to the relationship of those sets.
- Anticipating the consolidated linear score for an individual object of a set of variables utilizing said variables as a part of another set.

Based on the three mentioned objectives, there are several applications for CCA and its extensions (Kernel CCA (KCCA): multiple sets CCA (mCCA), Nonlinear CCA—such as Biometrics, Pattern recognition, Text translation and analysis, Remote sensing data processing, Blind source separation, Image retrieval, and Image processing—and analysis.

Assuming a linear relation between variates and the normality of data, normal CCA was used in this research. The next chapter will present a research-oriented approach and, afterwards, how the database was created and CCA applied to the datasets.

Chapter 3 – RESEARCH APPROACH

Aviation safety and risk analysis have been studied from many different aspects, as seen in the previous chapter. In mathematical and statistical analysis approaches, flight events data have primarily been studied. Researchers—whose works were previously reviewed—have tried to improve safety by building mathematical models that elaborate on accident causes, while providing precautionary measures to reduce flight risks. That said, most of the databases used in such research are directly retrieved from agencies like FAA and NTSB and therefor missing data points from causal reports.

This study aims to examine how correlations between flight variables (i.e. flight information, weather, airport, aircraft, and pilot associated variables)—or Independent Variables (IVs)—and incident/accident variables—or Dependent Variables (DVs)—are affected by differing factors. This emphasis on correlative analysis is intended to incorporate not only the aforementioned factors but also demonstrate this approach's ability to yield more precise 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 yield qualitative results, as well as enable us to apply the findings to more advanced mathematical modelling, which could be used by a variety of aviation personnel, such as flight dispatchers and crew schedulers, to optimise flight risk analysis. 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.

This chapter will present the research hypothesis, the data used in the analysis, data limitations, and methodology used in the research.

3.1 Research Hypothesis and Questions

This research supposes that there may be a statistical correlation relationship between Independent Variables (IV) and Dependent Variables (DV) of flight accidents/incidents. Capturing such correlation and defining which variables have higher effect on DVs will possibly lead to the detection of such factors in the future and help resolve them before they lead to accidents.

In general IVs can be categorized into five broad groups based on relative data: (1) flight information, (2) weather information, (3) airport information, (4) aircraft information, and (5) pilot information. This research focuses on measuring the effect of each respective category on the correlation between independent variables and dependent variables of aviation events (i.e. level of aircraft damage, severity of injury or fatality, etc.). The central research questions are:

1. Is there any significant difference between correlations of each category? In other words, can any of these categories mathematically be proved to have a higher effect on the outcome of aviation events (accidents/incidents)?
2. How can the significance in the previous question help improve flight safety?
3. How can the combination of different factors affect the severity of flight accidents or incidents?

This research has answered these questions by analyzing data from aircraft accidents and incidents within the United States over a specific period of time.

3.2 Data

To obtain meaningful results, three main requirements should be fulfilled: first to establish a comprehensive and reliable database, which this research has done. 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 variables.

A careful selection of variables was crucial for this analysis. Data selection requirements and methods for data pre-processing, variable selection, and grouping will be described below. After building the database, a multivariate statistical method will be introduced and applied to reveal correlations among variables and identify the most significant.

3.3 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 is added; upon completion of the investigation, the initial findings are replaced by "a final description of the accident and its probable cause" (NTSB, 2014).

In order to download a report, one must specify relevant information and submit a query. Preparing a database for retrieval requires information such as the provision of time intervals, locations, and 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 3-1 shows the details of 10 queries and Table 3-2 presents a summary of the query selected for the study (#10), 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 event records, which comprise a sizable statistical population for data analysis.

Table 3-1 Details of different queries on NTSB database

Query Number:	Query time interval:	Location:	Aircraft category:	Operation type:	Investigation type:	Word string used in the query:	Total number of records:
1	01/01/2003 to 12/31/2013	USA	Airplane	All general aviation	Accident/Incident	-	12,362
2	01/01/2003 to 12/31/2013	USA	Airplane	All general aviation	Accident/Incident	Air Traffic Controller	307
3	01/01/2003 to 12/31/2013	USA	Airplane	All general aviation	Accident/Incident	Traffic Controller	315
4	01/01/2003 to 12/31/2013	USA	Airplane	All general aviation	Accident/Incident	ATC	490
5	01/01/2003 to 12/31/2013	USA	Airplane	Part 121:Air Carrier	Accident/Incident	-	547
6	01/01/2003 to 12/31/2013	USA	Airplane	Part 121:Air Carrier	Accident/Incident	Air Traffic Controller	24
7	01/01/2003 to 12/31/2013	USA	Airplane	Part 121:Air Carrier	Accident/Incident	Traffic Controller	24
8	01/01/2003 to 12/31/2013	USA	Airplane	Part 121:Air Carrier	Accident/Incident	ATC	102
9	01/01/2003 to 12/31/2013	USA	Airplane	All general aviation	Accident/Incident	Traffic Controller	315
10	01/01/2003 to 12/31/2013	USA	Airplane	Part 121:Air Carrier	Accident/Incident	Probable Cause	508

Table 3-2 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 raw database.

Said database was obtained and all corresponding PCRs downloaded. One example of PCR is provided in Appendix A. The database consists of rows and columns in which rows correspond to events and columns to variables/factors. The database and PCR reports form the basis for the process of data preparation and database enhancement.

3.4 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:

1. 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 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 concerning the magnitude of the event:

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

2. New Variables

This study includes several critical variables, like pilot information, that are not provided in the raw NTSB database but are provided, either explicitly or implicitly, in the detailed Probable Cause Reports (PCRs). Following the systems engineering principles, and based on advice from experts in the FAA and National Oceanic and Atmospheric Administration (NOAA), a set of requirements for new variables were prepared. Once the

variables had been chosen, individual PCRs were carefully examined to incorporate the new data and build a more comprehensive database.

Figure 3-1 shows the details of factors from the raw database and others collected from narrative PCRs. Data shown in green are those used in the final analysis, which will be discussed shortly.

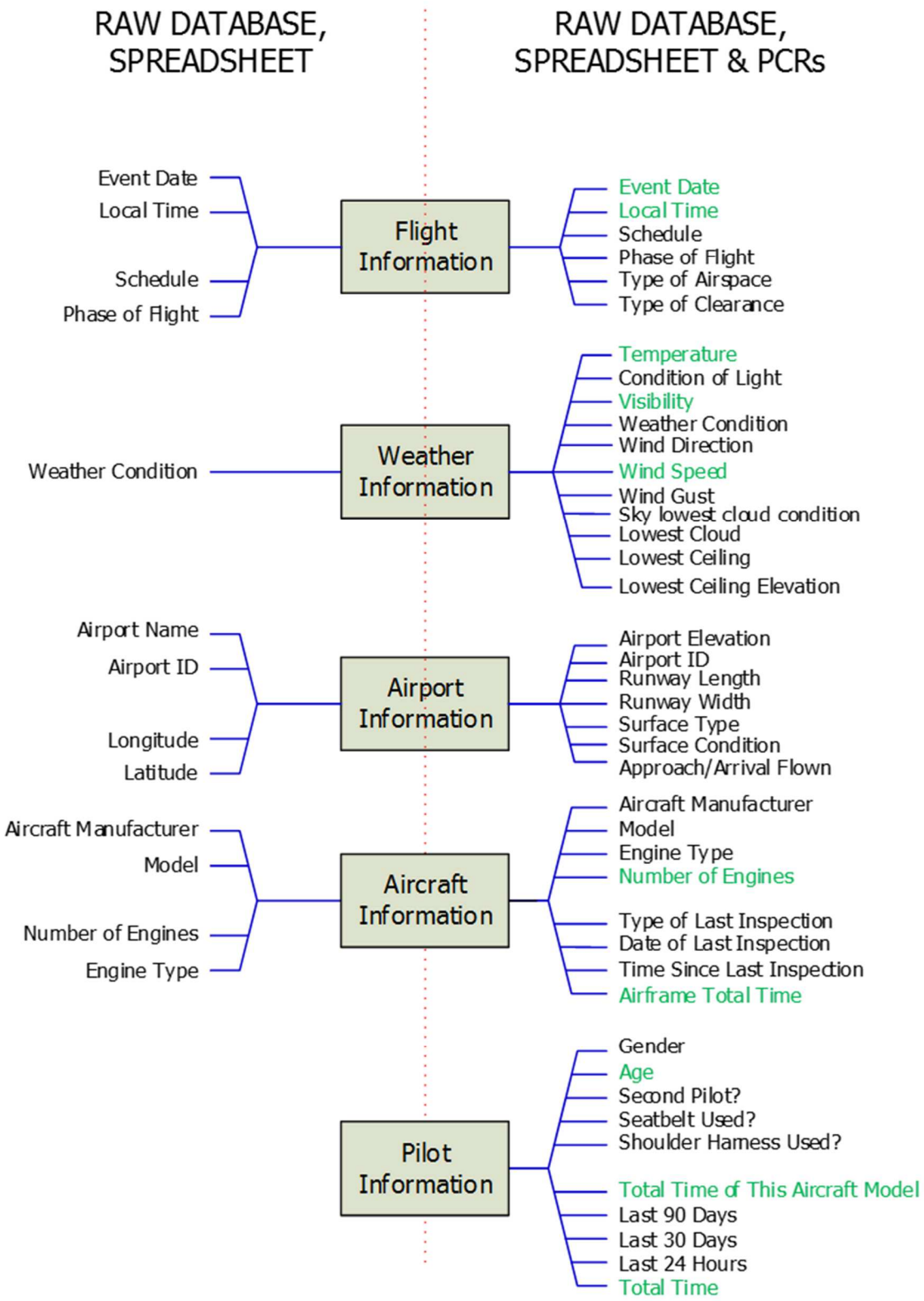

Figure 3-1 Database Improvement Using PCRs

3.5 Data from Additional Sources

Grouping and including these 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, 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—were consulted. These external sources filled critical gaps in the raw database. Database improvement efforts are depicted in Figure 3-2.

In some cases, the flight phase was not explicitly stated in the report, but was implicit in the narrative. In such cases, judgment was based on the flight phase of the PCR's narrative.

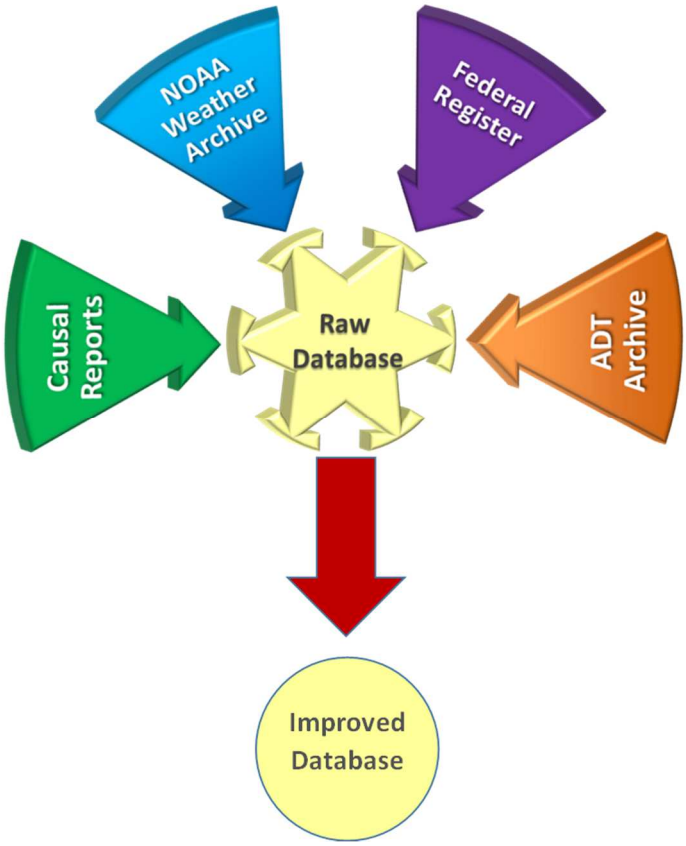

Figure 3-2 Database Improvement Using Additional Sources

3.6 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 irrelevant and insufficiently significant factors (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.

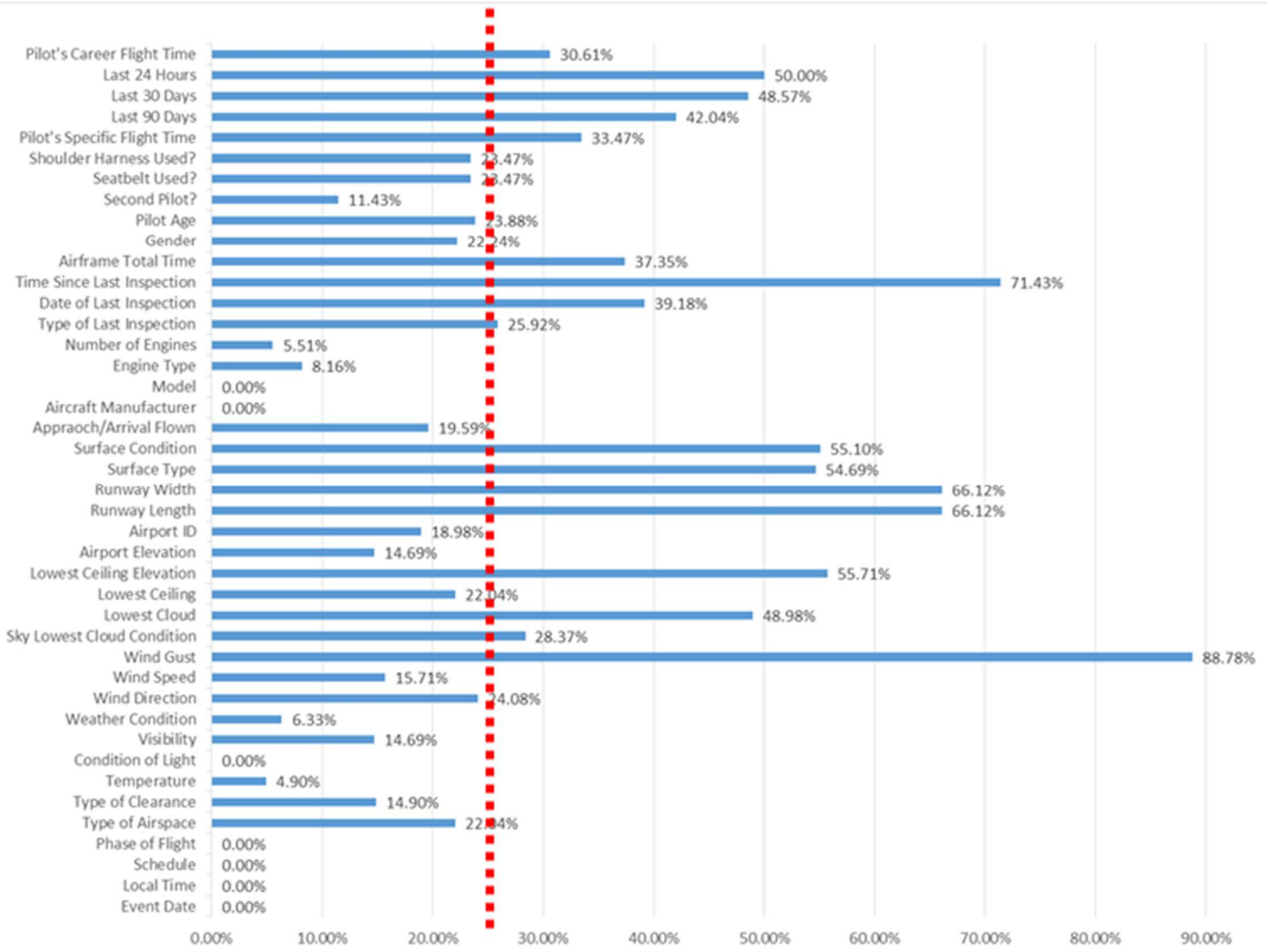

Figure 3-3 Percentage of Missing Values for Independent Variables

Percentages of missing values for each variable are shown in Figure 3-3, the red line representing 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 unquantifiable. To reveal their effects in the data analysis, however, they were included in the clustering phase, which will be discussed later.

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, logical assumptions were made. For example, wind speeds that were reported as 'calm' were assigned a numerical value of 0.5 mph.

After independent and dependent variables had been finalised, the database was ready to perform statistical analyses. Table 3-3 shows the resulting IVs and DVs, with information about type, range, and possible values for each variable.

Table 3-3 IVs and DVs for Statistical Analysis

	Independent Variables (IVs)	Type and Possible Values	Unit
1	Event date	Normalised number between 0 and 1	N/A
2	Event time	Normalised number between 0 and 1	N/A
3	Phase of Fight	Standing, Taxi, Take Off, Climb, Cruise, 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 Time (with accident/incident model)	Continuous values	Hour

	Dependent Variables (DVs)	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), Substantial (2), Destroyed (3)	N/A

3.5 Research Methodology

This study was conducted by performing analysis on a database retrieved from NTSB and improved, cleaned, and prepared as mentioned. CCA has been used as a statistical analysis tool, with its details and fundamentals previously outlined (Chapter 2). In applying it, correlations between DVs and IVs have been captured and multiple scenarios run. By monitoring the behavior of correlation while changing inputs of the model, we have tried to answer this research's central questions. The overall process of preparing

the database, performing multivariate statistical tests, and obtaining results is illustrated in Figure 3-4. The details will be presented in Chapter 4.

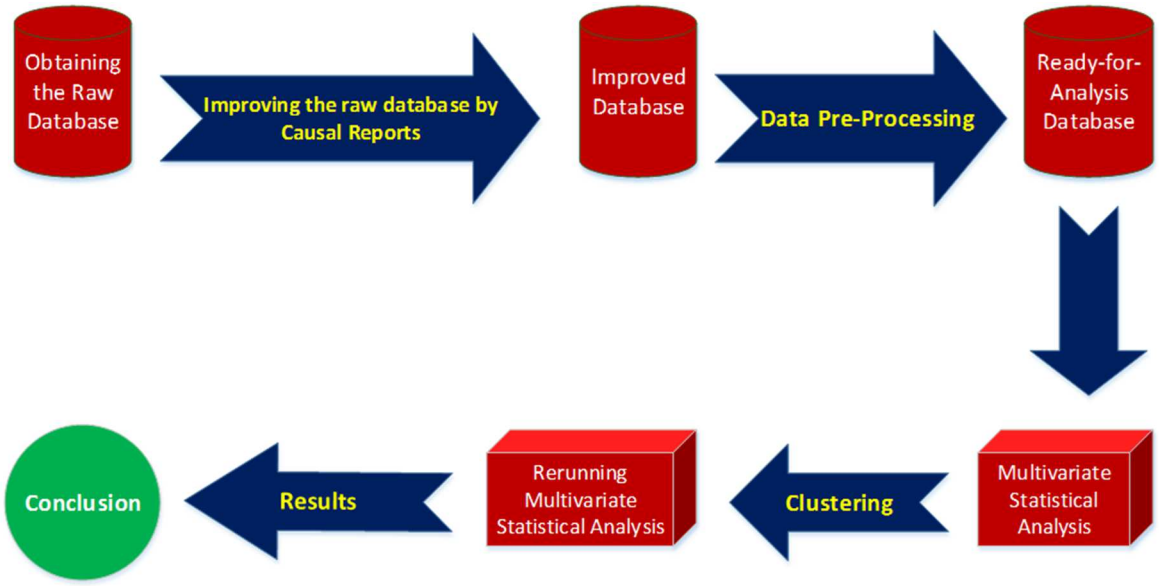

Figure 3-4 An Overview of Research Methodology

To better elaborate and propose a framework for similar studies, this process is summarized in the following step-by-step methodology:

Step 1 - Obtaining the raw database:

To begin any analysis or research, data collection is an essential step. Data can be retrieved either from existing databases or gathered using data collection methods such as questionnaire, interview, or direct observation. In our study we used the NTSB repository to obtain a raw database.

Step 2 – Database Improvement:

Before applying an analytical tool, we need to make sure that the database is as complete as possible; we need to identify any missing data and determine their values. As an example, if we have an independent variable called "temperature," and for some observations we couldn't obtain this value, we can refer to the "average daily temperature" database maintained by NOAA, find the value for the given day, and insert it into the database. In this research, we used other sources as demonstrated in chapter 3.2.2.

Step 3 – Selecting a proper data analysis tool

In this step, we need to decide on the data analysis tool, based on the nature of our research and the properties of our database. In the chapter 2.5 a brief introduction and literature review was presented to give researchers an idea about the different data analysis tools/models available. Because we were dealing with multi DVs and multi IVs, we selected CCA as the tool, assuming the linearity of our data.

Step 4 – Initial data analysis

Equipped now with a well-improved database and reasonable method, we can run an initial analysis. After the initial run, we will have to analyze our results. This will give us an idea as to whether or not our data modeling can provide reasonable results. If the data has noise or if there are outliers, these can be identified during this step. Also, we will be able to determine if the results are not as accurate as expected. If such is the case we will need to go to next step; otherwise, we will proceed to step 6.

Step 5 – Data analysis improvement

At this step, we need to improve our data analysis. This can be done by carefully studying the initial analysis and testing possible solutions. In our research, clustering was

proposed since the initial results were not promising. Applying clustering as a tool improved our data analysis however, providing much more meaningful results. Other possible techniques could also be pursued and applied during this step.

Step 6 – Discussion and conclusion

At this final step, we will need to discuss and explain the results of our analysis. This important step will provide the research's contribution. This discussion could be on the results themselves, explain the application of research, and/or provide a framework to be used by research stakeholders.

3.6 CCA and its Application

Prior research in other fields of science has demonstrated the uses and value of the CCA method to predict multiple DVs from multiple IVs (Bonner & Liu, 2005; Singh, Acharya, Mohanty, & Mishra, 2013; Singh, et al., 2012; Nourzad & Pradhan, 2016).

The aim with CCA is to identify and quantify the interrelations between a p-dimensional variable *X* and a q-dimensional variable *Y* (Dehon, Filzmoser, & Croux, 2000; Branco, Croux, Filzmoser, & Oliveira, 2005). The analysis looks for linear combinations of the original variables**,** $\mathbf{a^T X}$ and $\mathbf{b^T Y}$, that have maximal correlation.

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

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

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* (Nourzad & Pradhan, 2016). Each pair of canonical variates interprets the relationship in a given way. As discussed before, the strongest correlation between linear combinations of IVs and DVs will be captured using CCA method. The most significant pairs are those

with the highest correlations (Nourzad & Pradhan, 2016). The single variables that represent *X*-values and *Y*-values, respectively, are created using the formulas below:

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

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

In this research an approach was developed 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*'s and *V*'s (Nourzad & Pradhan, 2016). The model's effectiveness depends on how well the captured linear relationships fit. 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 3-5 depicts our model, in which the r-squared value will be measured and monitored depending on the change in the number of variables employed.

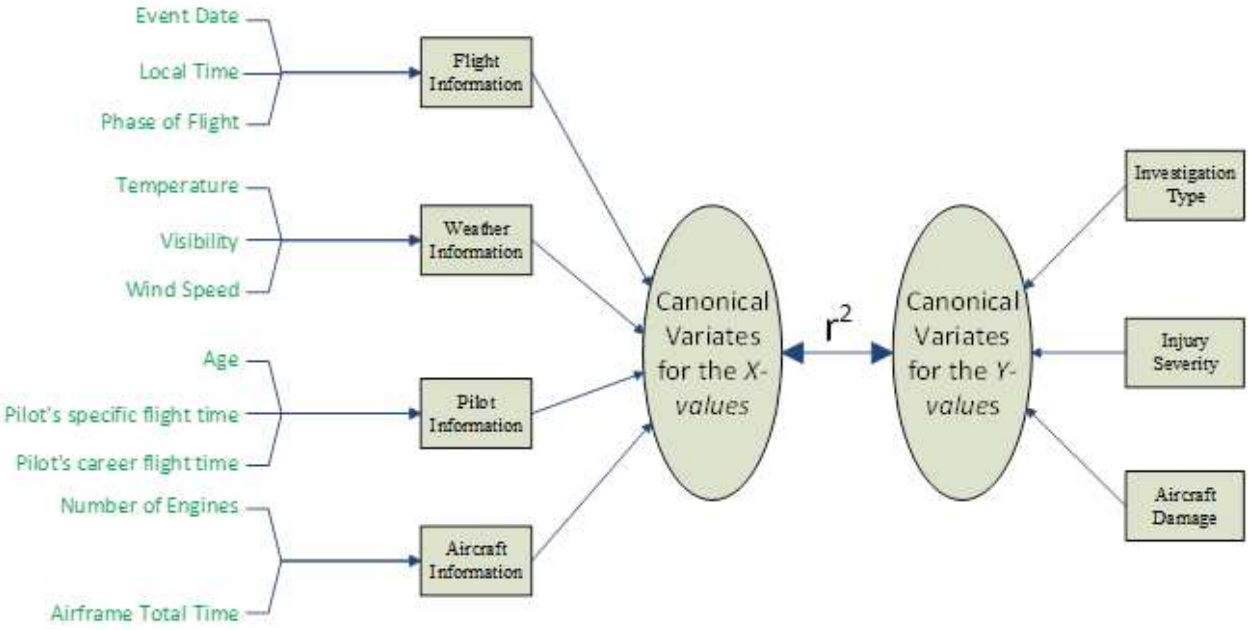

Figure 3-5 Research Model

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; however, classification is the underlying logic of both types of analysis. Homogeneity will be the basis for classification in both. Classification in respect to similarity in variables will shape clustering whereas factor analysis focuses on "homogeneity of variables resulting from the similarity of values assigned to variables" (Krebs, Berger, & Ferligoj, 2000).

Figure 3-6 is a generic illustration of how clustering can obtain stronger results from multivariate analysis.

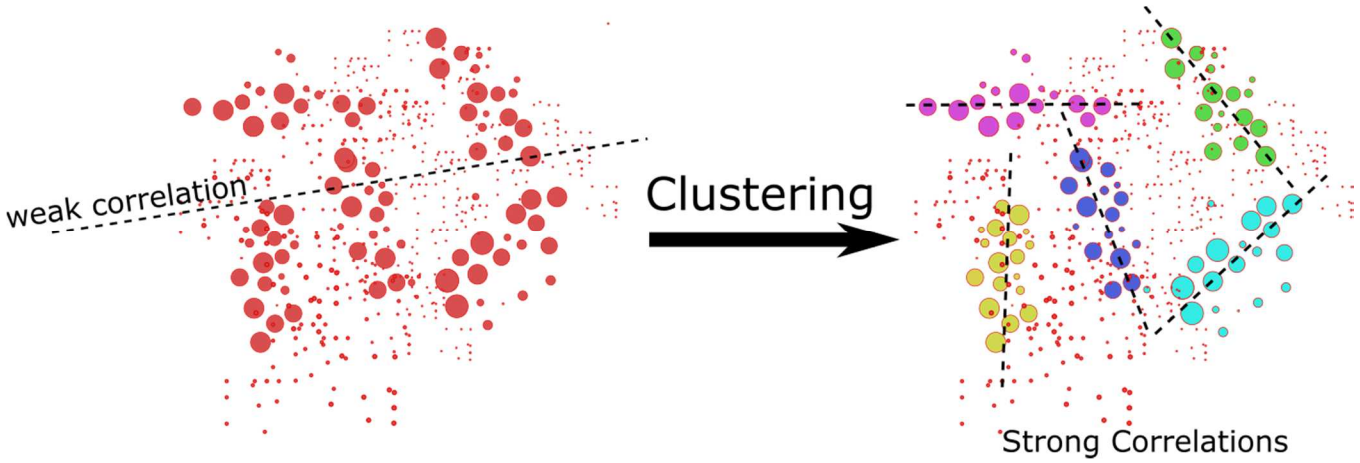

Figure 3-6 Data Clustering

To select the best variable to cluster, four variables capable of clustering 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. Figures 3-7 to 3-10 show the r-squared values for each cluster and while Figure 3-11 shows the average of r-squared values. As shown, clustering based on Phase of Flight yielded the highest r-squared values.

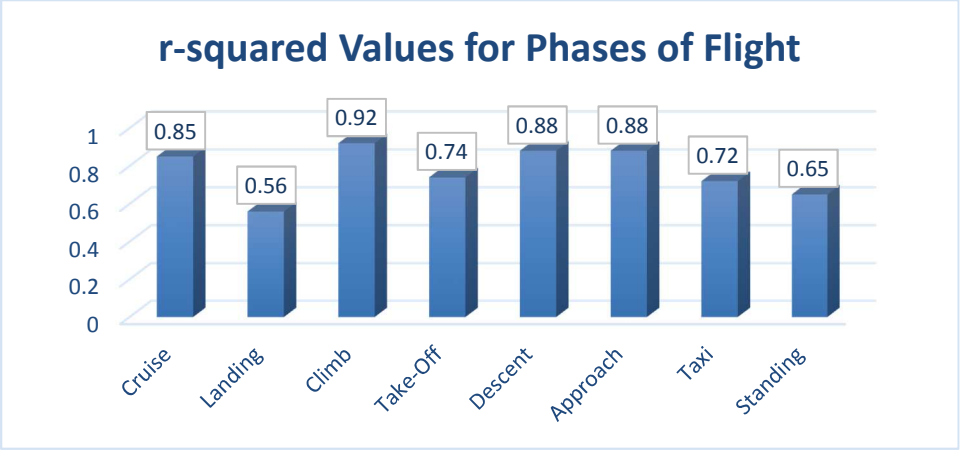

Figure 3-7 Correlation between DVs and IVs in Each Phase of Flight

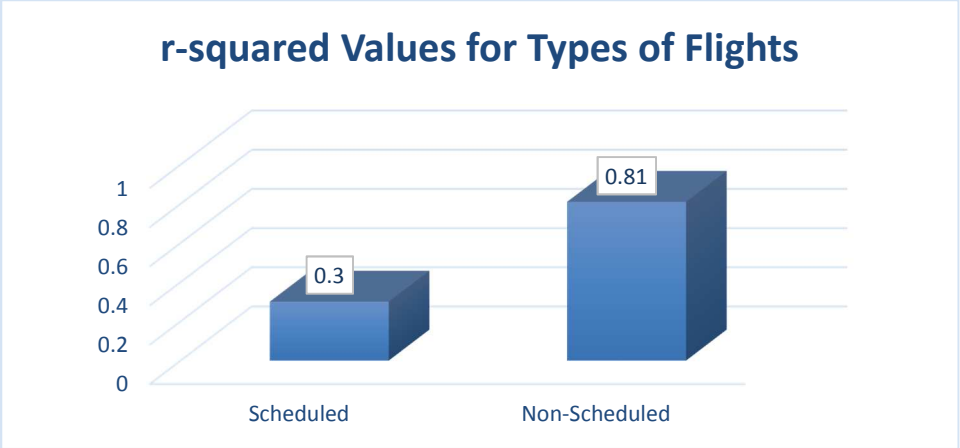

Figure 3-8 Correlation between DVs and IVs in Each Type of Flight

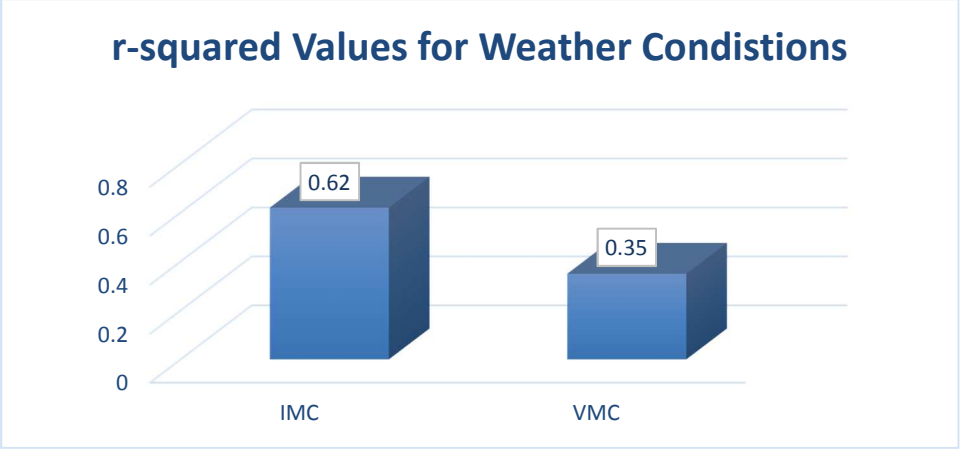

Figure 3-9 Correlation between DVs and IVs in Different Weather Conditions

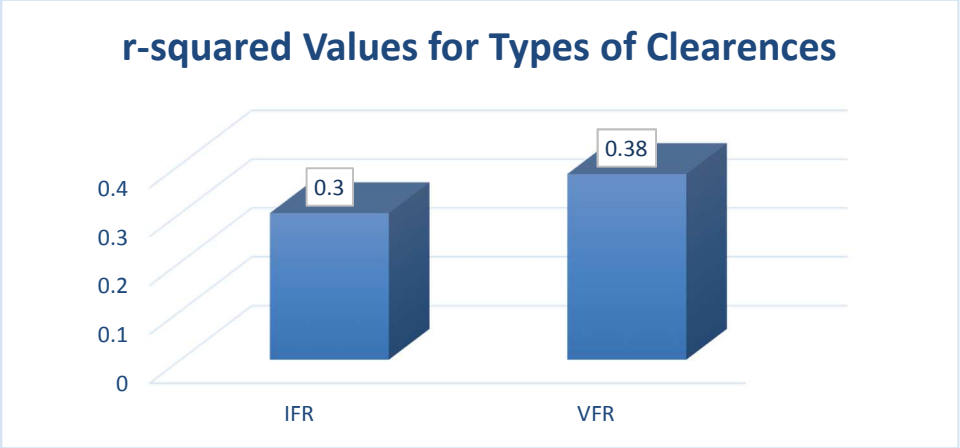

Figure 3-10 Correlation between DVs and IVs in Different Clearance Conditions

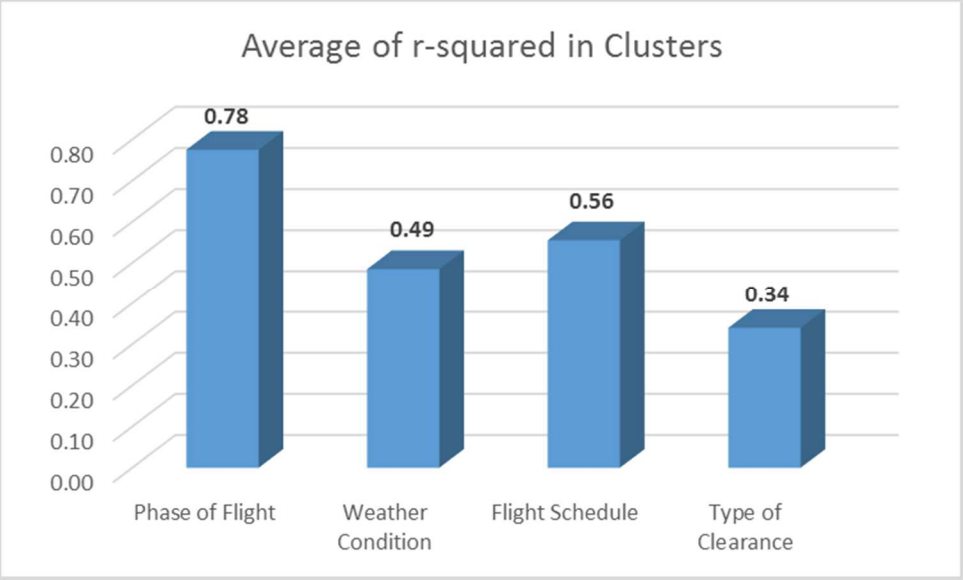

Figure 3-11 Average of r-squared Value for Different Clusters

As explained in Chapter 2, 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 3-11), 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 Chapter.

Chapter 4 – RESEARCH ANALYSES

4.1 Results

As mentioned, the first CCA run on the entire database did not yield fruitful results, the r-squared value showing a weak correlation. Following clustering, the correlations were strengthened significantly. Clustering was based on the Phase of Flight IV, which lent further relevance to non-numerical values. The r-squared values for all eight phases, with and without clustering, are shown in Figure 4-1.

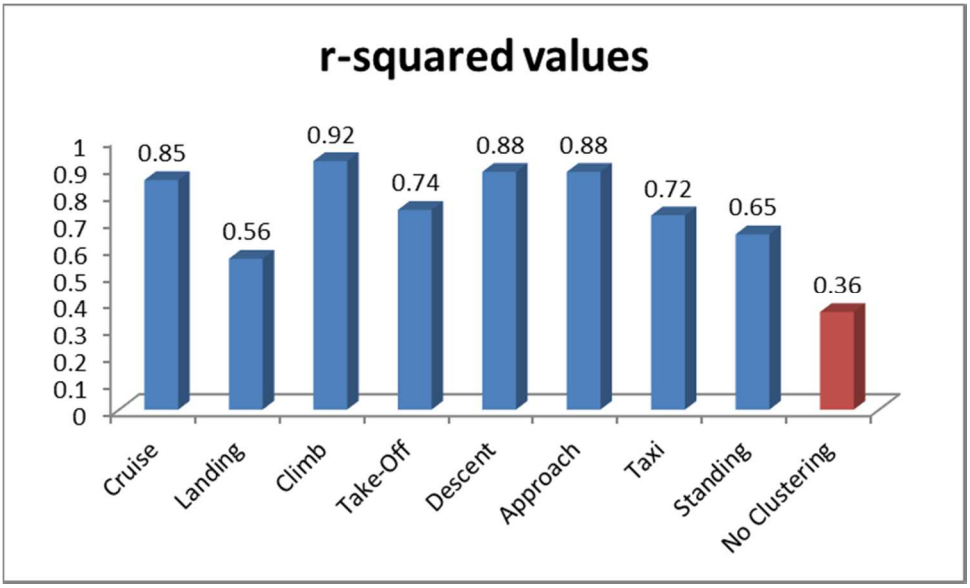

Figure 4-1 Significance of Clustering

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 4-1.

Table 4-1 r-squared Values for Different Run of CCA

		Cruise	Landing	Climb	Take-Off	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 the 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 4-2 and Figure 4-2.

Table 4-2 Drop in r-squared Values in %

	Cruise	Landing	Climb	Take-Off	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%

As shown in Table 4-2, 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.

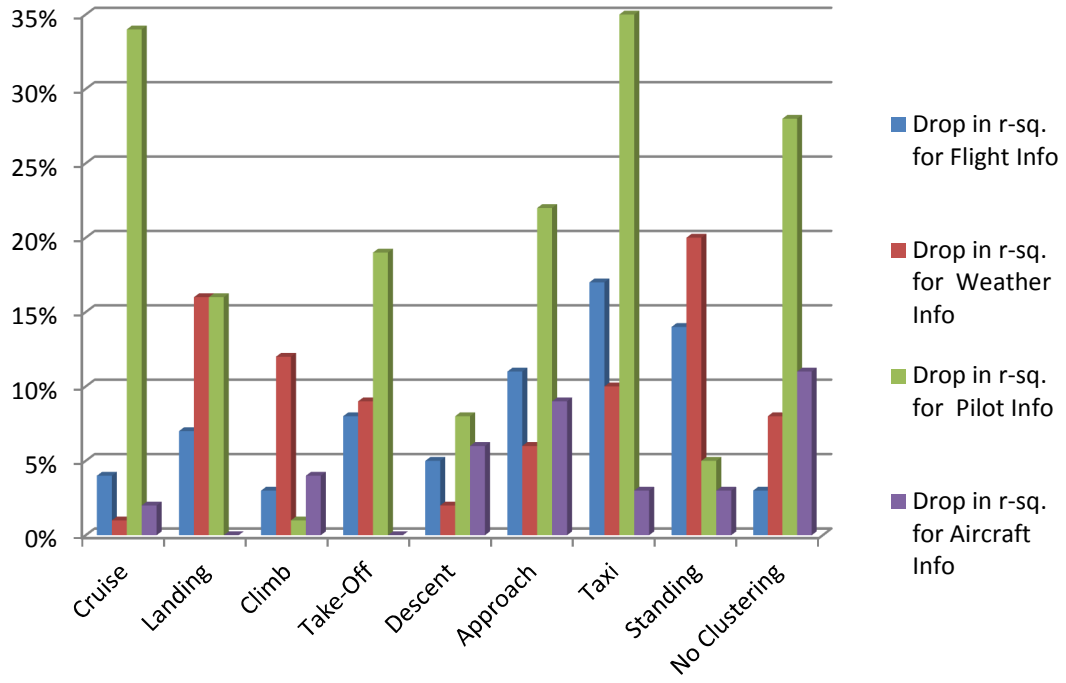

Figure 4-2 Drop in r-squared Values for Different Flight Phases

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 then that other factors, such as weather information or airport-related factors, would be more influential; this is corroborated by the results shown Figure 4-2. In the case of the Climb phase, the discrepancy in correlation may be attributed to the low number of events recorded during this phase.

4.2 Training Canonical Model – One Phase of Flight

So far we have been able to answer the first question of the research; we mathematically showed that pilot associated variables have the highest effect on the

correlation between IVs and DVs compared to other variables. Now we will continue building the model using CCA and calculating canonical variate vectors and other parameters in the equations and lastly finally verify the model.

To train the model, 88% of randomly selected data points from the cruise phase were used to build the model; the remaining 12% were reserved for testing the results. In this research, the number of variables in X and Y were 10 and 3. Therefore, three canonical variates were captured:

$$p = 10 \quad and \quad q = 3 \rightarrow K = \min(p, q) = 3$$

Then α and β calculated by MATLAB were retrieved from the software. These two matrices are canonical coefficients for the X (i.e., flight attributes) and Y (i.e., event outcome). The α and β matrices had two columns. Each column of α (and β) contained the coefficients of individual X_i (and Y_j) in the canonical variates.

$$\alpha = \begin{bmatrix} 1.05550624966061 & -2.09109385780267 \\ -1.63867045083864 & 1.83956643432116 \\ -0.0123600058134164 & 0.00134713792040443 \\ 0.0544241681549019 & 0.122243740504988 \\ -0.0246431064804647 & -0.113608337636036 \\ 1.55325008493533 & -0.0697293695293634 \\ -5.16913592982598E-06 & 5.63563075366726E-06 \\ 0.122930884117441 & -0.0431774888455296 \\ -0.000873171664976067 & -0.00009888441885426 \\ 0.000012555789836477 & 0.0002238248134536 07 \end{bmatrix}$$

$$\beta = \begin{bmatrix} -2.23644412354251 & 0.181108919970401 \\ -0.467529416013198 & 1.6521779889683 \end{bmatrix}$$

The two sets of canonical variates are the linear relationships between two linear combinations of X and two linear combinations of Y (i.e., $X.\alpha$ and $Y.\beta$). While infinite numbers of canonical variate sets can be defined, these canonical variates are the two sets of linear combinations that have the largest correlations among these infinite combinations.

To better illustrate these combinations we define U and V, the new variables as follows:

$$U = Y.\beta \text{ and } V = X.\alpha$$

The pairs resulting from this matrix multiplication can then be plotted. We plotted these two linear combinations in Figure 4-3 in which the x-axis shows a linear combination of the flight variables (V) and the y-axis shows a linear combinations of the event outcomes (U). The trend lines can be fitted by applying a linear regression analysis tool on the graphs. The equations corresponding to those lines are also shown on the graphs in Figure 4-3.

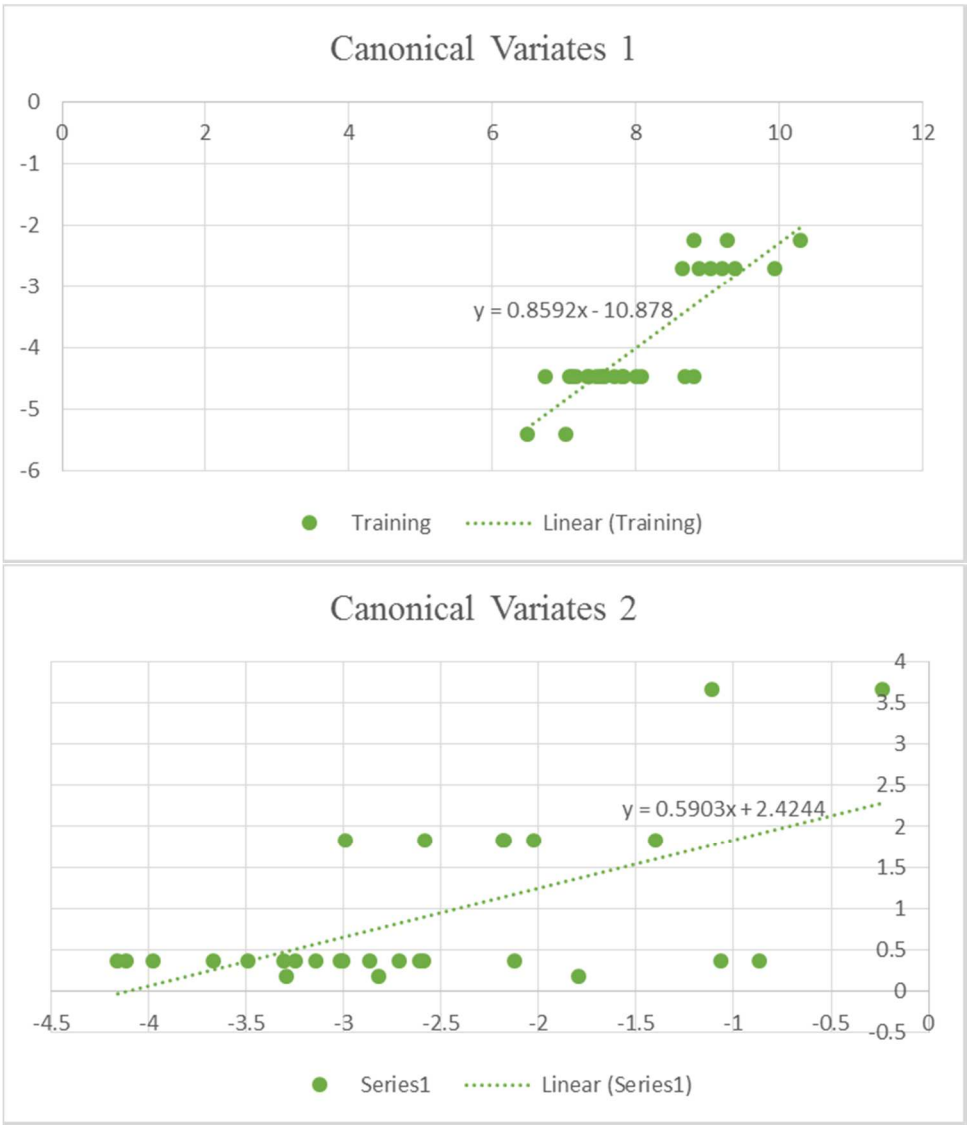

Figure 4-3 Canonical Variates 1 and 2

From these equations, the two new vectors, $\overrightarrow{c_1}$ and $\overrightarrow{c_2}$, can be created, which are the coefficients of the linearregressions (shown in Figure 4-3) as:

$$\mathbf{c_1} = \begin{vmatrix} 0.8592 \\ 0.5903 \end{vmatrix} \quad \& \quad \mathbf{c_2} = \begin{vmatrix} -10.878 \\ 2.4244 \end{vmatrix}$$

As we calculated vectors of coefficients of linear regressions ($\overrightarrow{c_1}$, $\overrightarrow{c_2}$), and matrices of canonical coefficients (α, β), so now for any flight with known factors (i.e., X_0), we can calculate:

$$C_0 = \overrightarrow{c_1}.(X_0.\alpha) + \overrightarrow{c_2}$$

Thus, the prediction model that predicts the event outcomes (i.e., Y_0) for a flight based on its attributes (i.e., X_0) is as follows:

$$Y_0 = C_0.\beta^{-1}$$

Although this model returns the possible event outcomes, it cannot completely replace the flight crew schedulers' experience or aviation regulations. This model can be used as supportive decision-making tool and should not be the only source for decision-making. Instead, this model can be used during the pre-planning process of numerous flight alternatives considering different pilots availability or weather conditions.

4.3 Model Verification – One Phase of Flight

To verify the model, 12% of reserved data points were used. We had to create our X_0 matrix as the input of the model. The arrays of this matrix are the IVs of those events, utilized as verifying points. For the sake of shortness and to avoid presenting too large a matrix below we present the first couple of rows of matrix X_0:

$$X_0 = \begin{bmatrix} 0.33 & 0.47 & 21 & 10 & 10 & 4 & 50293 & 55 & 848.0 & 7747.0 \\ 0.99 & 0.83 & -6 & 10 & 6.0 & 2 & 18350 & 52 & 15710 & 17210 \\ 0.68 & 0.62 & -25 & 0.0 & 5.8 & 2 & 45993 & 47 & 11900 & 17400 \\ 0.34 & 0.28 & 3.00 & 10 & 6.0 & 2 & 9495.00 & 37 & 2500 & 7500 \end{bmatrix}$$

Next we compute the matrix of C_0 using the previously presented formula:

$$C_0 = \overrightarrow{c_1}.(X_0.\alpha) + \overrightarrow{c_2}$$

All vectors and matrices on the right side of the equation have been calculated while building the model. As a result the C_0 is calculated:

$$C_0 = \begin{bmatrix} -0.26357654379188 & 1.24880319670248 \\ -3.65746986327213 & 0.381853220714256 \\ -4.96338768285533 & 1.05369027142854 \\ -4.22685988144223 & 1.58976156806589 \end{bmatrix}$$

The final step is multiplying this matrix with the inverse of β to calculate Y_0 and compare it with the values of DVs in order to calculate the model error. The corresponding Y_0 was calculated as follows:

$$Y_0 = \begin{bmatrix} -0.04 & 0.76 \\ 1.62 & 0.05 \\ 2.13 & 0.40 \\ 1.73 & 0.77 \end{bmatrix}$$

Excluding one data point that was an outlier, when comparing the arrays of Y_0 with the original DVs, it was observed that the results' "*average error*" was 11%. The error was calculated by comparing the model output (Y_0) to its corresponding Y_i from main database, then calculating the average error.

4.4 Training and Verification of Canonical Model: Other Phases of Flight

In the previous two sections, we built and verified a canonical model for events from the database of flights during their cruise phase. In those sections we presented a detailed method and matrices. To avoid repetition, in this section we will only present the results of training and verification CCA model for other phases of flight. Then in the next section, we will combine all data points together and build and verify the canonical model for all phases together.

For each phase of flight, we calculated necessary matrices in MATLAB and then computed vectors $\overrightarrow{c_1}, \overrightarrow{c_2}, and\ C_0$. Afterward, when the model was ready, we used reserved data points to verify the model. Percentages of training data points and verifying ones are presented in the Figure 4-4.

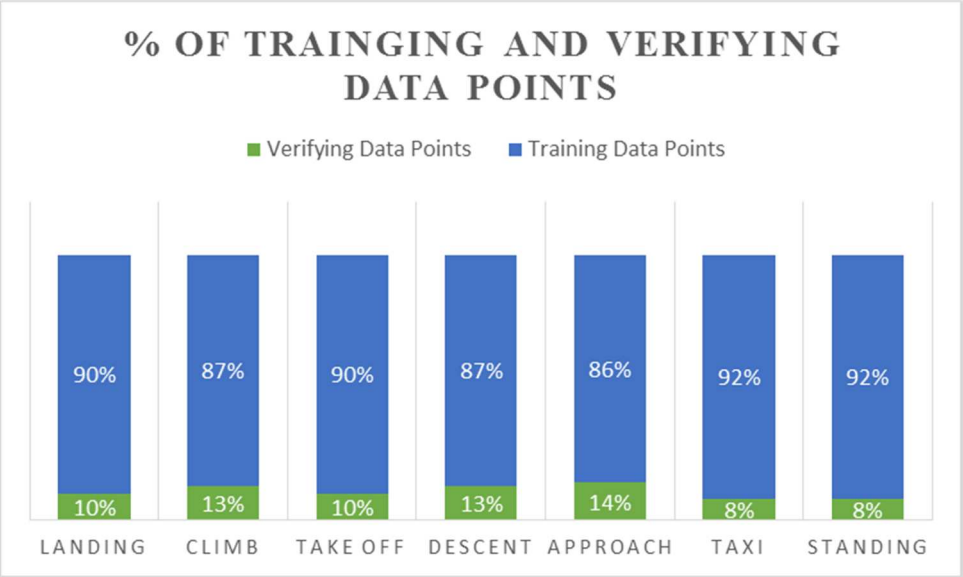

Figure 4-4 Percentages of Training and Verifying Data Points

By computing Y_0 matrices we were able to compare the values with their corresponding arrays from the main database. For two phases of flight (i.e. take-off and landing) the linear model failed to give meaningful results. For the other phases, the average errors from those comparisons are summarized in the Table 4-3.

Table 4-3 Summary of Average Error in Canonical Model

Phase of Flight	Average Error
Climb	3%
Descent	9%
Approach	14%
Taxi	8%
Standing	12%

4.5 Training Canonical Model – All Phases of Flight

In previous sections we built and verified a canonical model for events from the database based on the phase of flight. Now we will extend that model by including all events from all phases of flight to build a comprehensive model. To train the model 89% of randomly selected data points were used. Similar to the previous model, α and β were calculated and the model was built consequently.

$$\alpha = \begin{bmatrix} 0.082442 & -0.54377 & 1.423882 \\ -0.41506 & 2.715063 & 0.277226 \\ -0.01836 & -0.01307 & 0.02181 \\ -0.00599 & 0.06798 & 0.028038 \\ -0.03745 & 0.028324 & 0.06171 \\ 1.861494 & 0.754956 & -0.24289 \\ -1.3E-06 & 8.75E-06 & -4E-06 \\ 0.007043 & 0.026671 & -0.04426 \\ 1.01E-05 & 7.97E-05 & 0.000211 \\ -0.00012 & -1.6E-05 & -8.3E-05 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 0.179367 & 1.724885 & 1.340601 \\ -0.00837 & -0.08903 & 0.099587 \\ 1.096598 & -0.19024 & -0.46416 \end{bmatrix}$$

Similarly two matrices of *U* and *V* can be calculated and the pairs of (U_i,V_i) plotted. By plotting them and fitting trend lines we obtain these three equations:

$$y = 0.3782x + 0.715$$

$$y = 0.2086x + 1.3085$$

$$y = 0.1372x + 2.0077$$

As a result vectors of $\overrightarrow{c_1}$ and $\overrightarrow{c_2}$, can be created as follow:

$$\overrightarrow{c_1} = \begin{vmatrix} 0.715 \\ 1.3085 \\ 2.0077 \end{vmatrix} \quad \& \quad \overrightarrow{c_2} = \begin{vmatrix} 0.3782 \\ 0.2086 \\ 0.1372 \end{vmatrix}$$

Now the model is complete and we can verify it using reserved data points.

4.6 Model Verification – All Phases of Flight

To verify the model, 11% of reversed data points were used. The random combination of those events is shown in Figure 4-5.

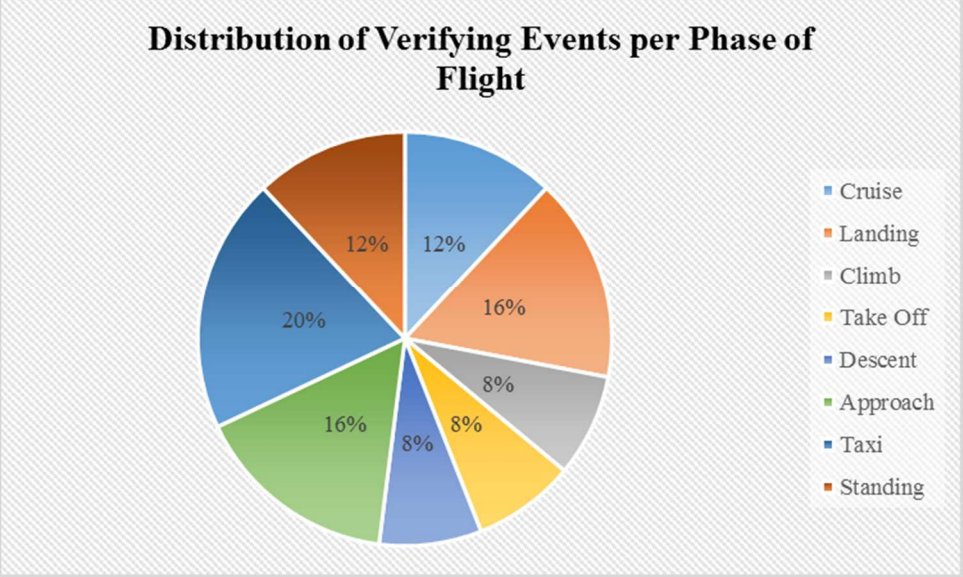

Figure 4-5 Combination of Testing Events

Similar to previous section processes the C_0 was computed enabling us to calculate the Y_0 from the X_0.

$$C_0 = \begin{bmatrix} 1.002 & 2.474 & 2.098 \\ 2.9185 & 2.697 & 1.718 \\ 1.4524 & 2.338 & 1.837 \\ 2.4544 & 2.316 & 1.682 \\ 0.8892 & 2.62 & 2.023 \\ 1.4865 & 2.363 & 2.191 \\ 1.324 & 2.36 & 1.909 \\ 1.698 & 2.313 & 2.027 \\ 1.6734 & 2.13 & 1.956 \\ 1.1881 & 2.617 & 1.987 \\ 1.6402 & 2.333 & 2.099 \\ 1.5933 & 2.344 & 1.86 \\ 1.5035 & 2.331 & 1.885 \\ 1.3257 & 2.456 & 1.965 \\ 1.2961 & 2.489 & 1.902 \\ 0.9609 & 2.446 & 1.839 \\ 1.7439 & 2.468 & 1.833 \\ 1.1176 & 2.41 & 1.625 \\ 0.9023 & 2.302 & 1.778 \\ 0.9486 & 2.361 & 1.974 \\ 1.4802 & 2.167 & 1.746 \\ 1.7909 & 2.018 & 1.965 \\ 2.2651 & 2.673 & 1.884 \\ 1.4944 & 2.529 & 1.983 \\ 1.0162 & 2.539 & 2.027 \end{bmatrix}$$

By continuing the verification process, the Y_0 was calculated and compared with the original *Y* matrix from testing events. The average error for this model for first, second, and third DVs are 9%, 23%, and 19% respectively with a total average of 17%.

4.7 Discussion

As we discussed in Chapter 3, to answer the main question of the research, firstly we had to have a reliable, thorough, and comprehensive database to be able to run the analysis. The raw database from NTSB was improved by incorporating data from probable cause reports prepared and published by the NTSB. As one of the most important contributions of this research, a great deal of time and effort was spent creating the database. However,

due to lack of certain data points, we had to drop some events from the database as well as some of the IV columns in the analysis. Due to the importance of that data, it is highly recommended to NTSB or FAA analysts to spend more effort to make sure missing data in the reports are minimal.

The current improved database is the result of three main processes. The first step was grouping the variables of the database in a more meaningful and useable order. The independent variables were grouped into 5 categories (flight information, weather information, airport information, aircraft information, and pilot information). The second step was to improve the database by introducing more variables in each category. The values for those variables had to be extracted from narrative reports corresponding to each event. This time-consuming step enriched the original database with important variables that were crucial for the canonical analysis. The narrative reports were also missing some values for variables that were necessary for the analysis. To have a comprehensive database we referred to other sources of data such as NOAA weather history to fill in the blanks. For example by having date, time, and location of an event we were able to search weather archives to retrieve temperature, wind speed, wind gust, and visibility of said time and date. This process was repeated for each single missing data point per each accident/incident. This eventually resulted in having a database that was not existing before, helping us in our research, and potentially helping other researchers in the future.

By consulting with experts and using their judgment, sets of DVs and IVs were developed and CCA was used to attempt to capture a meaningful correlation between them.

The first run of CCA on the improved database did not yield promising results, motivating the authors to find another way to tackle the problem. Amongst different

options, clustering the database based on the phase of flight significantly changed the results. The first interpretation of that change could be the importance of the phase of flight. Variables are better correlated if we consider each phase of flight separately. That said, in Chapter 4 we showed that in almost all phases of the flight pilot related factors have the highest level of effect on the correlation between dependent and independent variables. In the climb and standing phases of flight, we observed the lowest r-squared values. The lower level of correlation in the standing phase might be attributed to the pilot's low level of involvement; logically other factors, such as weather information or airport-related factors would be more influential, this is corroborated by the results as well. In the case of the Climb phase, the discrepancy in correlation may be attributed to the low number of events recorded during this phase. Having more events available gives statistical tools a better chance to capture relation and calculate correlation between DVs and IVs.

Our third and final step was to build an analytical model; we attempted to build a canonical model for each phase of flight using linear CCA. A total of 9 models were built and verified. Since the behavior of the data was unknown, we used regular CCA to capture linear relationships. As presented in section 4.4, meaningful models could be built for most phases of flight. However, linear modeling failed in providing meaningful results for only two instances—takeoff and landing phases. This could be due to the complexity of those phases of flight; an assumption of linear behavior of variables may not be a valid assumption. Firstly we would need to incorporate more data such as air traffic controllers' related data or navigation systems status; and secondly, use nonlinear modeling tools and techniques to see if better a model and results can be obtained. That said, the simple model in this research was able still to describe the database with average errors as presented in

Table 4-3. The model for the entire database led to higher average error (17%). This was because from the first phase of our research we observed a lower r-squared value while running CCA on the entire database without clustering (Figure 4-1). The results we presented in this chapter proved the significance of database preparation, improving the statistical tool's effectiveness to extract better results.

Chapter 5 CONCLUSION

In this dissertation, CCA was used to analyse an enhanced aviation events database prepared by the authors in order 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 (i.e. the NTSB's PCRs). Database enhancement was performed by studying all associated PCRs and retrieving new variables. The enhancement process included grouping, introduction of new variables, obtaining data from additional sources, and database cleaning. That said, this research was limited to events within the USA and as 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; this was in order to investigate the variables' effects on r-squared values between DVs and IVs.

A methodology was developed in this research that can be widely used in other research such as transportation safety research or similar analytical studies. The steps of this methodology are summarized in Table 5-1.

Table 5-1 Research Methodology Summary

Methodology Steps	**Application in this research**
Step 1 - Obtaining the raw database	Raw database retrieved from NTSB repository
Step 2 – Database Improvement	New variables, PCR data, using external data sources
Step 3 – Selecting a proper data analysis tool	Canonical Correlation Analysis
Step 4 – Initial data analysis	Weak correlation between DVs and IVs (0.36)
Step 5 – Data analysis improvement	Clustering
Step 6 – Discussion and conclusion	Identifying more significant variables, building descriptive model

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. These group factors are listed in Table 5-2 in the order of their significance as result of data analysis:

Table 5-2 Group Factors in the Order of Their Significance

Significance Order	Title of Group	Related Variables
First	Pilot related factors	- Age - Career flight time - Flight time with the aircraft model involved in the event
Second	Weather related factors	- Temperature - Visibility - Wind Speed
Third	Flight related factors	- Event Date - Event Time
Fourth	Aircraft related factors	- Number of Engines - Airframe Total Time

To conclude the research, several linear models were built for different phases of the flight as well as for the entire database. Linear models showed promising results for most phases of the flight. However, for phases with more non-linearity involved they failed to provide meaningful results. These models could be used by different stakeholders—such as risk managers, airline planners, crew schedulers, and dispatchers—to minimise flight risk and improve flight safety.

This research provides a framework for further inquiry and the construction of more reliable predictive models using the database we have made available. 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 to entering raw data in regular flight scheduling process, the above-mentioned predictive model could be used to assess the combination of those factors and the level of risk they impose. In the flight risk management process, the results of this research can be used as risk precursor analysis. This model can potentially tell schedulers that in specific weather conditions how assigning a low experience pilot would increase flight risk. 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-hour 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.

Because CCA was applied in this research, it imposes its limitations and assumptions. Linear relationship was assumed for all variables in each set and also between sets. Applying none-linear methods could improve results and contribute to the findings of our study. Widening the event 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 or better results might be obtained.

REFERENCES

(2016, 04 05). Retrieved from Secret Flying: http://www.secretflying.com/2016/04/planes-collide-at-indonesian-airport/

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 17-21.

Barnett, A., & Wang, A. (2000, April). Passenger-mortality risk estimates provide perspectives about airline safety. *Flight Safety Digest, 19*(4), pp. 1-12.

Boeing. (2015). *Statistical Summary of Commercial Jet Airplane Accidents, Worldwide Operations | 1959 – 2014.* Seattle, Washington: Boeing Commercial Airplanes. Retrieved from http://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf

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,.

Borga, M. (2001, January 12). *Canonical Correlation.* Retrieved from http://www.imt.liu.se/~magnus/cca/

Branco, J. A., Croux, C., Filzmoser, P., & Oliveira, M. R. (2005). Robust Canonical Correlations: A Comparative Study. *Computational Statistics, 20*(2), 203-229.

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).

Courville, T., & Thompson, B. (2001, April). Use of Structure Coefficients in Published Multiple Regression Articles: β is not Enough. *Educational and Psychological Measurement, 61*(2), 229-248.

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.

Czerwinski, D., & Barnett, A. (2004). *Airlines as baseball players: An alternative approach to evaluating air carrier safety record.* Atlantic City, NJ. USA.: Draft report prepared for FAA Tech Center.

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).

Dehon, C., Filzmoser, P., & Croux, C. (2000). Robust Methods for Canonical Correlation Analysis. *Data Analysis, Classification, and Related Methods*, 321-326.

Fan, X., & Konold, T. R. (2010). *Canonical Correlation Analysis. In G.R. Hancock, & R.O. Mueller (Eds.), Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers.* New York: Routledge.

Forest Resources Association. (2002, October). The Domino Safety Theory. *Loss Control Overview*. Rockville, Maryland: FOREST RESOURCES ASSOCIATION INC.

French, A., Chess, S., & Santos, S. (2005, June 02). *Canonical Correlation & Principal Components Analysis* . Retrieved from San Francisco State University: http://userwww.sfsu.edu/efc/classes/biol710/pca/CCandPCA2.htm

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.

Gellman Research Associates. (1997). *Safety reports: aviation safety data accessibility study index.* Washington, DC.: Federal Aviation Administration.

Graham, J. (2008). The general linear model as structural equation modeling. *J. Educ. Behav. Stat.*

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(12):2639–2664.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004, December). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation, 16*, 2639-2664.
Hawkins, F. (1987). *Human Factors in Flight (2nd Ed.)*. Aldershot, UK: Ashgate.
Hawkins, F. H. (1987). *Human factors in flight.* Aldershot, UK: Ashgate.
Henson, R. K. (1999). Multivariate normality: What is it and how is it assessed? *Advances in social science methodology*, 193-211.
Henson, R. K. (2002). The logic and interpretation of structure coefficients in multivariate general linear model analyses. *annual meeting of the American Educational Research Association.* New Orleans, LA.
History.com Staff. (2010). *9/11 ATTACKS*. Retrieved from History.com: http://www.history.com/topics/9-11-attacks
Holmström, M., & Clemesha, B. R. (2008). *Planetary Environments.* Committee on Space Research.
ICAO. (2001). *Aircraft accident and incident investigation, Annex 13 to the Convention on International Civil Aviation, ninth edition.* Montreal, Canada: International Civil Aviation Organization.
ICAO. (2012). *Safety Management Manual (SMM).* International Civil Aviation Organization.
James Reason HF Model. (2013, May 31). Retrieved from Skybrary Aviation Safety : http://www.skybrary.aero/index.php/James_Reason_HF_Model
Janic, M. (2000). An Assessment of Risk and Safety in Civil Aviation. *Journal of Air Transport Management, 6*, 43-50.
Janic, M. (2000). An Assessment of Risk and Safety in Civil Aviation. *Journal of Air Transport Management*, 43-50.
Jiao, J., Zhao, T., & Wang, W. (2009). Flight safety simulation system based on hybrid dynamic system. *Reliability and Maintainability Symposium, 2009. RAMS 2009. Annual* (pp. 236 - 241). Fort Worth, TX: IEEE.
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.*
Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis (sixth edition)*. New Jersey: Pearson Prentice Hall.
Kashef, M. (2016). An Assessment of Flight Variables Affecting Civil Aviation Accidents and Incidents. *Journal of Air Transport Studies*.
Knapp, T. R. (1978). . Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin*, 410-416.
Krebs, D., Berger, M., & Ferligoj, A. (2000). Approaching Achievement Motivation - Comparing Factor Analysis and Cluster Analysis. *New Approaches in Applied Statistics*.
Kuehl, R. O. (1999). *Design of Experiments: Statistical Principles of Research Design and Analysis 2nd Edition.* Emeritus .
Lazarick, R. T. (1998). Airport Vulnerability Assessment - An Analytical Approach. *SPIE 3575, Enforcement and Security Technologies.*
Lee, H. B., Katz, G., & Restori , A. (2010). The Application of Canonical Correlation to Two-Dimensional Contingency Tables. *Tutorials in Quantitative Methods for Psychology*, 1-15.
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.
LIN, P.-H. (2011). Safety Management and Risk Modelling in Aviation. Delft, GA, The Netherlands: Next Generation Infrastructures Foundation.
Liou, J. J., Tzeng, G.-H., & Chang, H.-C. (2007). Airline Safety Measurement Using a Hybrid Model. *Journal of Air Transport Management, 13*, 243–249. doi:10.1016/j.jairtraman.2007.04.008
Lundberg, J., Rollenhagen, C., & Hollnagel, E. (2009). What-You-Look-For-Is-What-You-Find – The consequences of underlying accident models in eight accident investigation manuals. *Safety Science*, 1297 - 1311.
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.

Mahdavi Damghani, B. (2013). The Non-Misleading Value of Inferred Correlation: An Introduction to the Cointelation Model. *Wilmott Magazine*.
McFadden, K. L., & Towell, E. R. (1999). Aviation human factors: a framework for the new millennium. *Journal of Air Transport Management*, 177-184.
McGarigal, K., Cushman, S., & Stafford, S. (2000). *Multivariate Statistics for Wildlife and Ecology Research.* New York: Springer.
Mercier, A., & Soumis, F. (2007). An integrated aircraft routing, crew scheduling and flight retiming model. *Computers & Operations Research, 34*, 2251 - 2265.
Mkondiw, M., Jumbe, C. B., & Wiyo, K. A. (2013). Poverty–Lack of Access to Adequate Safe Water Nexus: Evidence from Rural Malawi. *African Development Review, 25*(4), 537–550.
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).
Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med J.*, 69-71.
Murphy, T. P., & Levendoski, R. J. (1989). *Accident/Incident Data Analysis Database Summaries - Vol. I* . Washington, D.C.: U.S. Department of Transportation Federal Aviation Administration .
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 the Research on Risk and Safety Modelling in Civil Aviation. *THIRD INTERNATIONAL CONFERENCE ON RESEARCH IN AIR TRANSPORTATION.* FAIRFAX, VA.
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. (2016). Computational Modeling of Networked Infrastructures: Macroscopic Multivariate Approach. *Journal of Computing in Civil Engineering*.
NTSB. (2014). *NTSB Aviation Accident Database & Synopses*. Retrieved from www.ntsb.gov: http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx
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.
Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 25-45.
Perezgonzalez, J., Lee , D., LI , H., & ZiZhang, N. (2011, April 17). *SHELL model of human factors*. Retrieved from Wikidot.com: http://wikiofscience.wikidot.com/technology:shell-model-of-human-factors
Reason, J. (1990). *Human error.* New York: Cambridge University Press.
Reinhart, R. (1996). *Basic Flight Physiology(2nd Ed.).* New York, USA: McGraw-Hill.
Rencher, A. C. (2002). *Methods of Multivariate Analysis.* Canada: John Wiley & Sons, Inc.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis.* John Wiley & Sons.
Rodgers, J. L., & Nicewander, A. W. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 56-65.
Roelen, A. (2008). *Causal Risk Models of Air Transport.* Amsterdam: IOS Press.
Rose, N. L. (1992). Fear of flying: economic analyses of airline safety. *Journal of Economic Perspectives, 6*, 75–94.
Salkind, N. J. (2010). Encyclopedia of Research Design. In *Encyclopedia of Research Design* (pp. 867-870). SAGE Publications.
Schmitt, D., & Gollnick, V. (2016). *The Air Transport System.* Springer.
Senders, J. W., & Moray, N. P. (1991). *Human Error: Cause, Prediction, and Reduction.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.
Sherry, A., & Henson, R. K. (2005). Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. *JOURNAL OF PERSONALITY ASSESSMENT*, 37-48.
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.
Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences.* Mahwah: Erlbaum.
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
Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics.* New York: HarperCollins.
Tamasi, G., & Demichela, M. (2011). Risk Assessment Techniques for Civil Aviation Security. *Reliability EngineeringandSystemSafety, 96*, 892-899.
Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development*, 80-95.
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
Warner, R. (2008). *Applied Statistics: From Bivariate through Multivariate Techniques.* Thousand Oaks, CA: Sage.
Weenink, D. (2003). CANONICAL CORRELATION ANALYSIS. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, (pp. 81-89).
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 iterativeapproachtorobustandintegratedaircraftroutingandcrewscheduling. *Computers &OperationsResearch, 37*, 833 - 844.
Wiegmann, D. A., & Shappell, S. A. (2001). Human Error Perspectives in Aviation. *International Journal of Aviation Psychology, 11*(4), 341-357.
Wiegmann, D. A., & Shappell, S. A. (2003). *A Human Error Approach to Aviation Accident Analysis.* Ashgate Publishing Ltd.

Wu, S., Hrudey, S., French, S., Bedford, T., Soane, E., & Pollard, S. (2009). A role for human reliability analysis (HRA) in preventing drinking water incidents and securing safe drinking water. *Water Research*, 3227-3238.

Appendix A – Sample of PCR

This space for binding

National Transportation Safety Board FACTUAL REPORT AVIATION	NTSB ID: DCA10IA015	Aircraft Registration Number: N483A
	Occurrence Date: 12/13/2009	Most Critical Injury: None
	Occurrence Type: Incident	Investigated By: NTSB

Location/Time

Nearest City/Place	State	Zip Code	Local Time	Time Zone	
Charlotte	NC		2246	EST	

Airport Proximity: On Airport/Airstrip	Distance From Landing Facility:

Aircraft Information Summary

Aircraft Manufacturer	Model/Series	Type of Aircraft
BOEING	MD/82	Airplane

Revenue Sightseeing Flight: No Air Medical Transport Flight: No

Narrative

Brief narrative statement of facts, conditions and circumstances pertinent to the accident/incident:

*** Note: NTSB investigators used data provided by various sources and may not have traveled in support of this investigation to prepare this aircraft incident report. ***

HISTORY OF FLIGHT

On December 13, at 2246 Eastern Standard Time, a Boeing MD-82, N483A, scraped its right wingtip while landing on runway 36 Center at Charlotte Douglas International Airport (CLT), Charlotte, NC. The flight was registered to and operated by American Airlines under the provisions of 14 Code of Federal Regulations Part 121 as a passenger flight. Instrument flight rules (IFR) conditions prevailed for the flight, which operated on an instrument flight rules flight plan. The flight originated at Dallas-Fort Worth International Airport (DFW), Texas at 2035 Eastern Standard Time.

The DFW to CLT leg was flown by the first officer. There were no mechanical issues listed in the logbook of the airplane. The captain reviewed the CLT weather and briefed the first officer on the likelihood of flying a Category III (Cat III)approach upon arrival into CLT.

Departure from DFW was normal, and both pilots ate while discussing the weather in CLT and the fact that the captain would likely be required to conduct the approach based upon American Airlines' policy. American Airlines requires the captain to conduct all Cat III approaches, and when the weather is below 4000' runway visual range or 3/4 mile visibility. At 2130 EST the dispatcher sent flight 1402 the weather and Notices to Airmen. The CLT surface weather observation valid for the arrival time of the flight called for calm winds, visibility of 1/8 statute mile, a 100 foot ceiling, a temperature and dew point of 7 degrees Celsuus, and a barometric pressure of 30.08 inches of mercury. and a deg point of

The crew continued to check the CLT weather, and the captain decided to brief the Cat III approaches to both runway 36R and 36C. The captain said he chose the Cat III approach briefing as a "worst case scenario". According to the first officer, they conducted these briefings referencing several manuals, and completed the briefing several hundred miles from CLT. The first officer also told investigators that the captain briefed him that he expected them to visually pick up the airport early, and he would disconnect the autopilot and land visually.

The captain stated that the last time he had performed a CAT III approach was 9 months previously in the simulator, and had not performed a Cat III approach during line operations in "10-12 years" and "he was a little concerned about it." The first officer stated that he had not flown a Cat III approach in line operations in the last "14-15 years".

Following the briefing, the first officer conducted the descent into CLT. Some time past the SHINE intersection on the arrival to CLT, the flight crew switched pilot flying (PF) duties from the first officer to the captain by switching autopilot and autothrottle controls to the captain.

This space for binding

Narrative (Continued)

The first officer became the pilot monitoring (PM)and remained so for the rest of the flight.

When the airplane was established on final approach, the first officer said he checked for stable approach criteria and handled ATC communications. He said that ATC was busy launching airplanes and there was a lot of talking on the radio. The captain stated he believed the tower controller was clearing another airplane for takeoff while his flight was inside the outer marker, was concerned about it, but did not query ATC. The last reported visibility received by the flight crew was for an RVR of 1400' for runway 36C. There was no turbulence along the approach, and the crew stated the ride was "smooth" and normal.

The airplane was fully configured prior to arriving at the outer marker with the gear down, speedbrakes armed, and the airspeed about 170 knots. They slowed to approach speed after the outer marker. At about 1000 to 1500 feet on the approach, the first officer noticed a "momentary vibration" of the localizer needle, but there was no movement of the flight controls and no annunciator warnings, and he did not mention the needle movement to the captain. The captain saw the localizer "twitching" just before receiving a landing clearance from the tower at about 600 feet, but it was within limits according to the captain.

As the airplane descended through 300 feet above the ground, the captain began to acquire the runway environment visually, and noticed that the airplane was slightly right of the centerline of the runway with the nose pointed to the left and the airplane drifting toward the left. The captain disconnected the autopilot at that time and began flying the airplane manually. The bank angle aural alert responded as the airplane was being hand flown by the captain. The first officer looked inside the airplane to verify their altitude at about 200 feet, then looked out and noticed the airplane was then left of centerline and maneuvering with "a good amount of bank" to the right. The bank angle aural alert sounded again at about "30-32 feet" while the airplane was in a right correction back to the centerline. The airplane touched down approximately 1000 to 1500 feet down the runway to the left of centerline with the right wing tip coming into contact with the runway surface, and the left main wheels touching down in the grass adjacent to the runway.

At gate arrival, the captain entered "no items" in the maintenance logbook while the first officer began a post flight walk-around of the airplane. During his exterior inspection of the airplane, the first officer noticed minor damage to the right wing tip, right slat and a landing light. He returned to the cockpit and advised the captain. The captain made a logbook entry reflecting the "right wingtip touch" on landing. Neither pilot advised ATC of the incident.

FLIGHT CREW INFORMATION

The captain and first officer had flown together numerous times prior to this pairing. The incident flight was the fourth flight of the day, and occurred on the first day of a four day trip paring.

Both pilots had the proper flying and medical ratings for the incident flight. And, both pilots had successfully completed their last training and proficiency checks. A review of the Accident/Incident Data System and Enforcement Information System revealed no prior accident, incident or enforcement actions for either pilot. On December 14, 2009, the day after the incident, the flight crew was administered a drug and alcohol screening by an American Airlines contract laboratory. According to American Airline's Manager of Flight Safety, both test results for both pilots were negative.

The captain's flight times are as follows:

Total pilot flying time	15,793 hours
Total flying time last 24 hours	7 hours
Total flying time last 30 days	73 hours
Total flying time last 90 days	159 hours

The first officer's flight times are as follows:

Narrative (Continued)

Total pilot flying time 11,251 hours
Total flying time last 24 hours 7 hours
Total flying time last 30 days 52 hours
Total flying time last 90 days 188 hours

AIRCRAFT INFORMATION

According to post-incident interviews with investigators, neither pilot observed a failure in the autoflight computer system, the yaw damper system, the pitch control system or the flight director system at the time of the incident. The first officer stated that he successfully performed a Cat III autoland ground test on both sides of the DFGS prior to their departure from DFW to CLT.

Regarding the instrument landing system (ILS) ground transmitters, the CLT Tower Facility log revealed that between the hours of 2100 EST and 2400 EST on December 13, 2009, there were 21 arrivals to runway 36C at CLT. Entries into the log showed no PIREPs (Pilot Reports) or notations indicating anomalies with the ILS for runway 36C at CLT.

Copies of the FAA Technical Performance Records (FAA Form 8750-24 "LOC-Normal Radiated Parameters" and FAA Form 8750-23 "LOC-Normal Ground Check") covering the dates July 21, 2009 to December 18, 2009 for the CLT Runway 36C showed no abnormalities in the signal for the ILS 36C.

A search of the ASRS database covering reports for the CLT airport using "CLT", "CLT.AIPORT", "CLT.TOWER", "CLT.TRACON', and "ILS 36C" as search field criteria yielded no safety reports submitted from January 2005 to February 2010 concerning the ILS 36C at CLT.

ORGANIZATIONAL AND MANAGEMENT INFORMATION

During his post-incident interview, the captain stated that he disconnected the autopilot "sometime below 300 feet," and manually flew the airplane to landing.

American Airlines Airplane Operating Manual, Volume 1, page 40.1, stated that all Cat III approaches are to be flown by the autopilot and to an autoland. In interviews with the MD80 Fleet Training Manager, the MD80 Fleet captain, and an MD80 APD (Aircrew Program Designee), each stated that American Airlines trains its MD80 crews to autoland all Cat III approaches, rather than manually flying the airplane during the final descent to the runway.

Cat III Autoland Restrictions

American Airlines Flight Manual Part 1, Section 10, page 31, 7.1, "Use of Autoland" states, in part that autolands may be conducted to CAT I / II / III designated runways provided that the procedure does not have a chart note or NOTAM which renders the localizer unusable inside the runways threshold. The 11-5A Jeppesen approach chart for the ILS 36C approach to CLT had a note listed on the chart face that stated: IDQG LOC 36C unusable for rollout guidance.

The captain stated in his interview with investigators that he briefed a Cat III approach to both runways 36C and 36R at CLT. Title 14 Code of Federal Regulations (CFR) Part 121.603(a) states, in part that before beginning a flight, each pilot in command shall obtain all available current reports or information on airport conditions and irregularities of navigation facilities that may affect the safety of the flight. After the incident, on December 17, 2009, American Airlines issued an F473 message to append all flight plans and flight releases sent to all MD-80, Boeing 757, 767 and 777 aircraft with the following message statement (in part):

Subject: FM Part II Caution Re: Localizer Rollout Guidance Notes

This space for binding

Narrative (Continued)

AA Aircraft are not authorized to autoland on a runway where the following note is published in the notes section, as a ball note, or anywhere else on the approach chart: Localizer unusable for rollout guidance. AA Aircraft authorized to perform Cat II manual landing are also authorized to conduct Cat II manual landings on runways whose approach charts contain the reference note, provided that all FM-1 and airplane Vol. 1 operating manual requirements are satisfied.

Practice Autolandings and Cat III Practice Approaches

American Airlines Flight Manual Part 1, Section 10, page 31, 7.1 "Use of Autoland or 737 CAT III HUD" states in part:

A. Captains qualified on airplane types having autoland or CAT III HUD capability should make every effort to execute at least one such landing each month in the course of line operations in order to maintain familiarity and confidence with the system. These autolandings or HUD approaches should be made after considering factors including scheduled arrival time, traffic, etc. When conducting an autoland (including VMC) all CAT III crew coordination procedures must be followed to ensure all tracking and monitoring requirements are satisfied.

The captain stated in his post-incident interview that "he hadn't flown an autoland in 10-12 years, or one in the sim [simulator] in 9 months," and further added that "he didn't feel comfortable with it". The first officer stated in his interview that, while crews would see a Cat II and Cat III in recurrent simulator training "every 9 months", he personally had not flown a Cat II or Cat III approach in line operations in "14-15 years". The MD80 Fleet Captain told investigators that they [American Airlines] encouraged crews to do FCCs for their own confidence in the autoland system of the airplane, but there was no requirement to do it. Following this incident, the MD80 Fleet Training Manager sent internal company bulletins to the MD80 crews reminding them to conduct Cat III approaches "to remain current".

Go Around Policy

American Airlines had a no-fault go-around policy. Pilots were told to execute every approach with the presumption that a missed approach was a successful outcome, and asked to plan each approach through the missed-approach procedure and make the decision to land only when all criteria are safely satisfied.

According to the AA Flight Manual, Section 10 "Approaches and Landing," the captain "should give every consideration to a recommendation by another cockpit crewmember that a missed approach be executed." The American Airlines MD80 Fleet Captain told investigators that the first officer could also call for a go around, and did not have to explain why.

During post-incident interviews, the first officer stated that during the approach, neither pilot mentioned a go around. The captain stated in his interview that he did not consider a go around.

Conditions Requiring a Missed Approach

According to the AA DC-9 Operating Manual, Section 40.6 "Approach-Landing-Go-Around", the following conditions would require a go around from a Cat III Approach:

1. Below 300 feet (RA), if satisfactory tracking performance was not maintained.

2. Failure of required airplane or ground equipment prior to DH (Cat II) or prior to touchdown (Cat III).

Approaches and Landings

The MD80 at American Airlines was authorized to conduct approach and landing operations to Category III minimums.

This space for binding

Narrative (Continued)

The lowest RVR value authorized for the fail passive autoland system on the MD80 was 600 feet RVR. Cat III approaches could be flown to an altitude of 50 RA. This altitude was considered a decision height (DH) at which the captain (pilot flying on a Cat III approach) would make a decision to allow the airplane to land automatically or go around, based on whether or not he/she had the Touchdown Zone or the Touchdown Zone Lights in sight.

The American Airlines DC-9 Operating Manual, Section 40.6 "Approach - Landing - Go-Around" stated that when the airplane arrived at the decision height (DH), the first officer (Pilot Monitoring) would call out "minimums". The captain callouts would be:

1. If the required visual reference for-a successful landing is satisfied:
Call out - "Landing"

2. If the required visual reference for a successful landing is not satisfied or subsequently lost:

Call out - "Go-around" and execute missed approach.

American Airlines DC-9 Operating Manual, Section 20.1 regarding standard callouts stated that the pilot flying would make the following callouts:

"Landing" or "Go Around" at DA, MDA, or DH as appropriate on instrument approaches.

During the incident, the first officer did not make the required "minimums" callout, and the captain did not make a "landing" or "go around" callout.

Stable Approaches Criteria

FAA Advisory Circular AC 120-71A "Standard Operating Procedures for Flight Deck Crewmembers" Appendix 2, states in part:

An approach is stabilized when all of the following criteria are maintained from 1000 feet height above touchdown (HAT), or 500 feet HAT in VMC, to landing in the touchdown zone:

1. The airplane is on the correct track.
2. After glide path intercept, or after the FAP, or after the derived fly-off point (per Jeppesen) the pilot flying requires no more than normal bracketing corrections to maintain the correct track and desired profile (3° descent angle, nominal) to landing within the touchdown zone.

A stabilized approach means the airplane must be:

- At Approach Speed (VREF + additives).
- On the proper flight path at the proper sink rate.
- At stabilized (spooled up) thrust.

These requirements must be maintained throughout the rest of the approach for it to be considered a stabilized approach. If the stabilized approach requirements cannot be satisfied by the minimum stabilized approach heights or maintained throughout the rest of the approach, a go-around was required.

The FAA Advisory Circular appendix goes on to state in part:

Normal bracketing corrections relate to bank angle, rate of descent, and power management.

This space for binding

Narrative (Continued)

Recommended ranges are as follows (operating limitations in the approved airplane flight manual must be observed, and may be more restrictive):

Bank angle: Maximum bank angle permissible during approach is specified in the approved operating manual used by the pilot, and is generally not more than 30°; the maximum bank angle permissible during landing may be considerably less than 30°, as specified in that manual.

Excessive Bank Angles

According to the American Airlines Flight Manual Supplement, Section VII.1, there were six modes of the Ground Proximity Warning function (GPWS). Mode 6 provided optional callouts for descent through predefined radio altitudes between 2,500 and 10 feet AGL and excessive roll or bank angle warning. Bank angle warning provided over-banking protection during approach, climb out, and cruise. Additionally, the warning protected against wing or engine strikes during the landing.

American Airlines DC-9 Operating Manual, Section 55.5 "Approach-Landing-Go-Around" stated in part:

With the main landing gear compressed the landing light will strike the runway at approximately eight and one-half degrees of bank. Excessive bank angles should be avoided at low altitudes.
Updated on Jun 27 2011 2:29PM

This space for binding

National Transportation Safety Board

FACTUAL REPORT AVIATION

NTSB ID: DCA10IA015
Occurrence Date: 12/13/2009
Occurrence Type: Incident

Landing Facility/Approach Information

Airport Name	Airport ID:	Airport Elevation	Runway Used	Runway Length	Runway Width
Charlotte Douglas Intl	CLT	748 Ft. MSL	36C	10000	300

Runway Surface Type: Asphalt

Runway Surface Condition: Dry

Approach/Arrival Flown: ILS

VFR Approach/Landing: None

Aircraft Information

Aircraft Manufacturer	Model/Series	Serial Number
BOEING	MD/82	

Airworthiness Certificate(s): Transport

Landing Gear Type: Tricycle

Amateur Built Acft?	Number of Seats:	Certified Max Gross Wt.	Number of Engines:
No		LBS	2

Engine Type:	Engine Manufacturer:	Model/Series:	Rated Power:

- Aircraft Inspection Information

Type of Last Inspection	Date of Last Inspection	Time Since Last Inspection	Airframe Total Time
		Hours	Hours

- Emergency Locator Transmitter (ELT) Information

ELT Installed?/Type	ELT Operated?	ELT Aided in Locating Accident Site?
Yes / Unknown	No	No

Owner/Operator Information

Registered Aircraft Owner	Street Address	City	State	Zip Code
American Airlines	4333 Amon Carter Bvd.	Fort Worth	TX	76155

Operator of Aircraft	Street Address	City	State	Zip Code
American Airlines	4333 Amon Carter Bvd.	Fort Worth	TX	76155

Operator Does Business As:

Operator Designator Code: AALA

- Type of U.S. Certificate(s) Held:

Air Carrier Operating Certificate(s): Flag Carrier/Domestic

Operating Certificate:

Operator Certificate:

Regulation Flight Conducted Under: Part 121: Air Carrier

Type of Flight Operation Conducted: Scheduled; Domestic; Passenger Only

This space for binding

National Transportation Safety Board

FACTUAL REPORT AVIATION

NTSB ID: DCA10IA015

Occurrence Date: 12/13/2009

Occurrence Type: Incident

First Pilot Information

Name	City	State	Date of Birth	Age

Sex:	Seat Occupied:	Occupational Pilot?	Certificate Number:

Certificate(s):

Airplane Rating(s):

Rotorcraft/Glider/LTA:

Instrument Rating(s):

Instructor Rating(s):

Current Biennial Flight Review?

Medical Cert.:	Medical Cert. Status:	Date of Last Medical Exam:

- Flight Time Matrix	All A/C	This Make and Model	Airplane Single Engine	Airplane Mult-Engine	Night	Instrument Actual	Instrument Simulated	Rotorcraft	Glider	Lighter Than Air
Total Time										
Pilot In Command(PIC)										
Instructor										
Instruction Received										
Last 90 Days										
Last 30 Days										
Last 24 Hours										

Seatbelt Used?	Shoulder Harness Used?	Toxicology Performed?	Second Pilot? Yes

Flight Plan/Itinerary

Type of Flight Plan Filed: IFR

Departure Point	State	Airport Identifier	Departure Time	Time Zone
Fort Worth	TX	DFW	2035	EST

Destination	State	Airport Identifier	
Same as Accident/Incident Location		CLT	

Type of Clearance: IFR

Type of Airspace:

Weather Information

Pilot's Source of Wx Information:

This space for binding

National Transportation Safety Board

FACTUAL REPORT
AVIATION

NTSB ID: DCA10IA015

Occurrence Date: 12/13/2009

Occurrence Type: Incident

Weather Information

WOF ID	Observation Time	Time Zone	WOF Elevation Ft. MSL	WOF Distance From Accident Site NM	Direction From Accident Site Deg. Mag.

Sky/Lowest Cloud Condition: Ft. AGL | Condition of Light: Night

Lowest Ceiling: Overcast 300 Ft. AGL | Visibility: SM | Altimeter: "Hg

Temperature: °C | Dew Point: °C | Weather Conditions at Accident Site: Instrument Conditions

Wind Direction: | Wind Speed: | Wind Gusts:

Visibility (RVR): 1400 Ft. | Visibility (RVV) SM

Precip and/or Obscuration:

Accident Information

Aircraft Damage: Minor | Aircraft Fire: None | Aircraft Explosion None

- Injury Summary Matrix	Fatal	Serious	Minor	None	TOTAL
First Pilot				1	1
Second Pilot				1	1
Student Pilot					
Flight Instructor					
Check Pilot					
Flight Engineer					
Cabin Attendants				3	3
Other Crew					
Passengers				110	110
- TOTAL ABOARD -				115	115
Other Ground					
- GRAND TOTAL -				115	115

This space for binding

National Transportation Safety Board

FACTUAL REPORT
AVIATION

NTSB ID: DCA10IA015

Occurrence Date: 12/13/2009

Occurrence Type: Incident

Administrative Information

Investigator-In-Charge (IIC)

Robert P. Benzon

Additional Persons Participating in This Accident/Incident Investigation:

Victoria E Anderson
Federal Aviation Administration
Washington, DC