



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING THE INCIDENCE OF MILITARY AVIATION SAFETY MISHAPS

by

Ryan C. Herrmann

June 2022

Thesis Advisor:
Second Reader:

Ruriko Yoshida
Jefferson Huang

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 2022	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE PREDICTING THE INCIDENCE OF MILITARY AVIATION SAFETY MISHAPS			5. FUNDING NUMBERS	
6. AUTHOR(S) Ryan C. Herrmann				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Air Force Safety Center, Kirtland Air Force Base, NM 87117			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) U.S. Air Force instruction requires commanders to conduct annual Mishap Prevention Analysis and develop safety plans that reverse adverse trends and guide prevention actions. Current analysis methods often fail to identify emerging trends in a timely manner and lack the ability to forecast the rate and trend of future mishaps based on historical data, resulting in arbitrary mishap rate reduction goals for the coming year. To properly safeguard service members and government property, commanders and safety staffs require timely trend identification and incidence forecasting provided by sound data analytic processes to target and combat safety trends. Using Air Force Safety Automated System (AFSAS) mishap and total flying hour data, we construct Time-Series Seasonal Decomposition and ARIMA predictive models that provide monthly trend, seasonality, and natural occurrence rates of Class A and B aviation mishaps and forecast the by-month mishap incidence rate for the following year. Each model provides valuable safety planning measures of performance. The seasonal model identifies months of peak incidence for safety plan targeting. The fluctuations of the ARIMA model establish the "Band of Normal" that identifies whether trends are exceeding natural rates. We develop a tool designed to be used by safety personnel at echelon to perform this analysis and predictive modeling, enabling them to focus on safety plan development instead of expending precious manhours processing data.				
14. SUBJECT TERMS military aviation, aviation safety, ARIMA, seasonal decomposition, regression, predictive model, incidence rate, AFSAS, data analysis, trend analysis, time series data			15. NUMBER OF PAGES 133	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**PREDICTING THE INCIDENCE OF MILITARY AVIATION SAFETY
MISHAPS**

Ryan C. Herrmann
Captain, United States Army
BS, United States Military Academy, 2013
MA, Webster University, 2017

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
June 2022**

Approved by: Ruriko Yoshida
Advisor

Jefferson Huang
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

U.S. Air Force instruction requires commanders to conduct annual Mishap Prevention Analysis and develop safety plans that reverse adverse trends and guide prevention actions. Current analysis methods often fail to identify emerging trends in a timely manner and lack the ability to forecast the rate and trend of future mishaps based on historical data, resulting in arbitrary mishap rate reduction goals for the coming year. To properly safeguard service members and government property, commanders and safety staffs require timely trend identification and incidence forecasting provided by sound data analytic processes to target and combat safety trends. Using Air Force Safety Automated System (AFSAS) mishap and total flying hour data, we construct Time-Series Seasonal Decomposition and ARIMA predictive models that provide monthly trend, seasonality, and natural occurrence rates of Class A and B aviation mishaps and forecast the by-month mishap incidence rate for the following year. Each model provides valuable safety planning measures of performance. The seasonal model identifies months of peak incidence for safety plan targeting. The fluctuations of the ARIMA model establish the “Band of Normal” that identifies whether trends are exceeding natural rates. We develop a tool designed to be used by safety personnel at echelon to perform this analysis and predictive modeling, enabling them to focus on safety plan development instead of expending precious manhours processing data.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Project Origins and Problem Specification	1
1.2	Moral Imperative	2
1.3	Ties to USAF Mishap Analysis Program	3
1.4	Research Questions	3
2	Background	5
2.1	Key Mishap Definitions.	5
2.2	Literature Review	8
2.3	Human Factor Analysis and Classification System	12
2.4	Privileged Safety Information	15
2.5	United States Air Force Safety Automated System Data	15
2.6	National Transportation Safety Board and Federal Aviation Administration Data for Proofs of Concept.	17
3	Methodology	21
3.1	Overview of Predictive Modeling Methodologies.	21
3.2	Proof of Concept 1: Simple Linear Regression.	28
3.3	Proof of Concept 2: Time Series Analysis	45
3.4	Applying Proof of Concept 2 to AFSAS Data	58
3.5	Final Model Selection and Use Cases	66
3.6	Normalization of Incidence Prediction	68
3.7	The Aviation Mishap Incidence Predictor Tool.	69
4	Findings	85
4.1	Number of Mishaps per Month	85
4.2	Mishap Rate per Total Flying Hours	92

5 Conclusion	101
5.1 Enumerated Answers to Research Questions	101
5.2 Recommendations for Future Research.	101
5.3 Conclusion.	103
 List of References	 105
 Initial Distribution List	 107

List of Figures

Figure 1.1	Military Aviation Mishap Losses between 2013 and 2018	2
Figure 2.1	Decision Algorithm Defining Whether a Mishap has Occurred . .	6
Figure 2.2	Decision Algorithm Defining Whether a Class A or Class B Mishap has Occurred	7
Figure 2.3	Flow Chart Depicting Error Back Propagation for Forecasting Avia- tion Mishap Rates	11
Figure 2.4	Swiss Cheese Model for Aviation Mishaps	13
Figure 2.5	Human Factor Analysis and Classification System (HFACS) Code Tree	14
Figure 3.1	Occurrence of Class A Mishaps in National Transportation Safety Board (NTSB) Data	31
Figure 3.2	Occurrence of Class B Mishaps in NTSB Data	31
Figure 3.3	Histogram of Federal Aviation Administration (FAA) Total Opera- tions by Year	32
Figure 3.4	Pairs Plot of proof of concept (POC)1 Data Indicating a Sparse Data Set with Multicollinearity	33
Figure 3.5	Performance Analysis of Initial Class A simple linear regression (SLR) Model	34
Figure 3.6	Fit of Initial Class A SLR Model to Actual Class A Occurrence .	34
Figure 3.7	Performance Analysis of Initial Class B SLR Model	35
Figure 3.8	Fit of Initial Class B SLR Model to Actual Class B Occurrence .	36
Figure 3.9	Performance Analysis of Transformed Class A SLR Model	37
Figure 3.10	Fit of Transformed Class A SLR Model to Actual Class A Occurrence	38

Figure 3.11	Performance Analysis of Transformed Class B SLR Model	39
Figure 3.12	Fit of Transformed Class B SLR Model to Actual Class B Occurrence	39
Figure 3.13	Performance Analysis of elastic net regularization (ENR)-Informed Class A SLR Model	40
Figure 3.14	Fit of ENR-Informed Class A SLR Model to Actual Class A Occurrence	41
Figure 3.15	Performance Analysis of ENR-Informed Class B SLR Model . .	42
Figure 3.16	Fit of ENR-Informed Class B SLR Model to Actual Class B Occurrence	42
Figure 3.17	90% Confidence and Prediction Intervals of Transformed Class A SLR Model	43
Figure 3.18	90% Confidence and Prediction Intervals of ENR-Informed Class A SLR Model	43
Figure 3.19	90% Confidence and Prediction Intervals of Transformed Class B SLR Model	43
Figure 3.20	90% Confidence and Prediction Intervals of ENR-Informed Class B SLR Model	43
Figure 3.21	NTSB Class A Mishap Seasonal Decomposition	47
Figure 3.22	NTSB Class A Mishap Seasonality and Trend	47
Figure 3.23	NTSB Class B Mishap Seasonal Decomposition	48
Figure 3.24	NTSB Class B Mishap Seasonality and Trend	48
Figure 3.25	NTSB Class A Mishap time series analysis (TSA) Model Performance under One-Step Forward Horizon Predictions	49
Figure 3.26	NTSB Class A TSA Fitted Models	50
Figure 3.27	NTSB Class B Mishap TSA Model Performance under One-Step Forward Horizon Predictions	50
Figure 3.28	NTSB Class B TSA Fitted Models	51

Figure 3.29	NTSB Class A Seasonal Decomposition Model Prediction	52
Figure 3.30	NTSB Class A Seasonal Decomposition Model Prediction	52
Figure 3.31	NTSB Class B Seasonal Decomposition Model	53
Figure 3.32	NTSB Class B Seasonal Decomposition Model Prediction	53
Figure 3.33	NTSB Class A Seasonal Auto Regressive Integrated Moving Average (ARIMA) Model Prediction	55
Figure 3.34	NTSB Class A Seasonal ARIMA Model Prediction	55
Figure 3.35	NTSB Class B Seasonal ARIMA Model Prediction	56
Figure 3.36	NTSB Class B Seasonal ARIMA Model Prediction	56
Figure 3.37	Air Force Safety Automated System (AFSAS) Class A Mishap Sea- sonal Decomposition	59
Figure 3.38	AFSAS Class A Mishap Seasonality and Trend	59
Figure 3.39	AFSAS Class B Mishap Seasonal Decomposition	60
Figure 3.40	AFSAS Class B Mishap Seasonality and Trend	60
Figure 3.41	AFSAS Class A Mishap TSA Model Performance under One-Step Forward Horizon Predictions	61
Figure 3.42	AFSAS Class A TSA Fitted Models	61
Figure 3.43	AFSAS Class B Mishap TSA Model Performance under One-Step Forward Horizon Predictions	62
Figure 3.44	AFSAS Class B TSA Fitted Models	62
Figure 3.45	AFSAS Class A Seasonal Decomposition Model Prediction . . .	63
Figure 3.46	AFSAS Class B Seasonal Decomposition Model Prediction . . .	64
Figure 3.47	AFSAS Class A Seasonal ARIMA Model Prediction	65
Figure 3.48	AFSAS Class B Seasonal ARIMA Model Prediction	65
Figure 3.49	Aviation Mishap Incidence Predictor (AMIP) User Input Window	72

Figure 3.50	AMIP Dashboard Interactive Graphics Mouse Hover Feature . . .	77
Figure 3.51	AMIP Dashboard Interactive Graphics Utility Buttons	77
Figure 3.52	AMIP Main Dashboard	78
Figure 3.53	AMIP Class A Dashboard	79
Figure 3.54	AMIP Class B Dashboard	79
Figure 3.55	AMIP Model Fit and Performance Dashboard	80
Figure 3.56	AMIP Class A Models for the T-38 MDS	82
Figure 3.57	AMIP Class B Models for the T-38 MDS	83
Figure 4.1	AFSAS Class A Mishap Seasonality — Number of Mishaps . . .	85
Figure 4.2	AFSAS Class A Mishap “Band of Normal” — Number of Mishaps	86
Figure 4.3	AFSAS Class B Mishap Seasonality — Number of Mishaps . . .	87
Figure 4.4	AFSAS Class B Mishap “Band of Normal” — Number of Mishaps	88
Figure 4.5	AFSAS Class A Mishap Seasonal Model Forecast — Number of Mishaps	89
Figure 4.6	Historical Trend in USAF Aviation Mishap Rates between 1950 and 2018	91
Figure 4.7	AFSAS Class A Mishap Seasonality — Mishap Rate	93
Figure 4.8	AFSAS Class A Mishap “Band of Normal” — Mishap Rate . . .	94
Figure 4.9	AFSAS Class B Mishap Seasonality — Mishap Rate	95
Figure 4.10	AFSAS Class B Mishap “Band of Normal” — Mishap Rate . . .	96
Figure 4.11	AFSAS Class A Mishap Model Forecasts	98

List of Tables

Table 4.1	Class A Seasonal Model Predictive Performance	89
Table 4.2	Comparison of Class A Seasonality to “Band of Normal”	97
Table 4.3	Comparison of Class B Seasonality to “Band of Normal”	97
Table 4.4	Class A Model Predictive Performance — Mishap Rate	99

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AFI	U.S. Air Force instruction
AFSEC	U.S. Air Force Safety Center
AFPC	U.S. Air Force Personnel Center
AFSAS	Air Force Safety Automated System
AIC	Akaike information criterion
AMIP	Aviation Mishap Incidence Predictor
ARIMA	Auto Regressive Integrated Moving Average
AWS	Amazon Web Services
BPNN	back propagation neural network
CY	calendar year
DAF	Department of the Air Force
DOD	Department of Defense
DODI	Department of Defense instruction
EDA	exploratory data analysis
ENR	elastic net regularization
FAA	Federal Aviation Administration
FY	fiscal year
HFACS	Human Factor Analysis and Classification System
ICAO	International Civil Aviation Organization

IFR	instrument flight rules
MAPE	mean absolute percentage error
MASE	mean absolute scaled error
MDS	mission design series
MOP	measure of performance
MSE	mean squared error
NAVSAFECEN	U.S. Navy Safety Center
NTSB	National Transportation Safety Board
NCMAS	2020 National Commission on Military Aviation Safety
NPS	Naval Postgraduate School
ORSA	Operations Research/Systems Analyst
POC	proof of concept
PSI	privileged safety information
RMSE	root mean squared error
SLR	simple linear regression
SMAPE	symmetric mean absolute percentage error
TSA	time series analysis
TFH	total flying hours
UAV	unmanned aerial vehicle
USA	U.S. Army
USAF	U.S. Air Force
USACRC	U.S. Army Combat Readiness Center

USMC	U.S. Marine Corps
USN	U.S. Navy
VFR	visual flight rules
VIF	variance inflation factor

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Military aviation safety programs are currently restricted to the rudimentary analysis of safety mishap data to identify emerging trends and incidence rate. This analysis, often conducted manually or utilizing unsuitable software solutions, fails to identify emerging trends in a timely manner and lacks the ability to accurately forecast the rate and trend of future aviation mishaps based on historical data. This rudimentary and untimely analysis results in arbitrary goals set by commanders to decrease the rate of mishaps for the next fiscal year (FY) or calendar year (CY). Using AFSAS mishap and total flying hour data provided by the U.S. Air Force Safety Center (AFSEC), we construct TSA Seasonal Decomposition and ARIMA predictive models that provide monthly trend, seasonal, and natural rates of occurrence for Class A and B military aviation mishaps and forecast the incidence rate by month for the following year.

This research will directly assist commanders and chiefs of safety in the performance of their duties. The U.S. Air Force (USAF) Mishap Prevention Program requires commanders and chiefs of safety know the types of mishaps and mishap rates for their command, and directs them to develop risk mitigation strategies based on sound mishap analysis (United States Air Force Safety Center 2020, p. 77). U.S. Air Force instruction (AFI) 91-202, Paragraph 5.4, specifies that this analysis must:

- develop specific actions to reverse adverse trends;
- identify successes, problem areas, and trends to measure safety program effectiveness and guide prevention actions; and
- be conducted annually (United States Air Force Safety Center 2020, p. 77).

The models and tool developed by this research are designed to directly assist in the development of realistic goals for annual safety programs by forecasting the next FY or CY of mishap incidence based on historical data. This research also provides commanders and chiefs of safety with indicators for when trends are exceeding normal rates of occurrence, as well as provides measures of performance (MOPs) to measure safety program effectiveness.

The Seasonal Decomposition model will assist commanders and chiefs of safety with identifying historical yearly seasonal trends by determining historical peaks and lows of

mishap incidence. Given that the predictions generated by this model for the next FY would incorporate historical seasonality, safety professionals would use the forecasted peaks and lows of the forecasted FY to design the annual mishap prevention plan, proactively targeting the peak month with the intent to reduce the likeliness of a mishap occurring from any of the top three causes of accidents. Safety professionals at echelon already know the leading causes of both Class A and B mishaps through use of the HFACS.

The ARIMA model will assist commanders and chiefs of Safety to identify their mishap rate “Band of Normal.” The “Band of Normal” establishes the band of what can be considered the normal or “natural” rate of mishap occurrence for each mishap class. The upper and lower limits of the fluctuations in the fitted model establish the upper and lower limits of this band. Predictions performed by the ARIMA model inform USAF commanders and chiefs of safety if the incidence is predicted to stay within, exceed, or fall below the natural rate of occurrence. The “Band of Normal” also provides a valuable measure of performance (MOP) for annual safety program performance:

- If the actual occurrence of mishaps trend above this band, commanders and chiefs of safety should analyze their mishaps for any emerging causal or contributory HFACS trends in an effort to curb additional mishap rate increases.
- If the actual occurrence of mishaps trend below this band, commanders and chiefs of safety should analyze their safety program for what actions reduced the rate of mishaps.
- If the actual occurrence of mishaps trend within the band, commanders and chiefs of safety should recognize that the rate of mishaps observed during this period occurred within the natural range of occurrence. However, this should not discourage commanders and chiefs of safety from pursuing the reduction of their mishap rates to the lower end of the band through proactive programs.

This use of the “Band of Normal” to detect whether mishap trends are exceeding normal rates is similar to the methodology utilized by International Civil Aviation Organization (ICAO) to detect mishap rate deviation (International Civil Aviation Organization 2022).

It must be noted that the ARIMA “Band of Normal” use case acknowledges an uncomfortable truth. A goal of zero mishaps was set by Gen (Ret) John P. Jumper, then USAF Chief of Staff, in December 2002, (Kreisher 2006) and in subsequent years, the goal of zero mishaps

has grown to become a mindset that permeates the USAF Safety Program. However, the “Band of Normal” recognizes that there is a rate of natural mishap occurrence that cannot be fully eliminated because of the dangerous and random nature of military aviation coupled with human performance factors.

Our research examined USAF mishap data provided by AFSEC from AFSAS. Analysis was conducted on more than ten years of data spanning from 01 October 2010 to 31 March 2021.

Final Findings for USAF-wide Class A Mishap Rates

Class A Final Findings		
Model	Low	High
Seasonality	1.14 (JAN21)	2.29 (MAY20)
Band of Normal	1.10	1.90

Peak mishap rate incidence occurs in May each year; low incidence occurs in January or March each year. This trend is constant across ten years of data. It is natural to have 1.10 to 1.90 Class A mishaps per 100,000 total flying hours (TFH) occur per month.

Final Findings for USAF-wide Class B Mishap Rates

Class B Final Findings		
Model	Low	High
Seasonality	2.06 (MAY20)	2.83 (JUN21)
Band of Normal	1.85	2.99

Mishap incidence spikes in June and July; incidence significantly decreases in August, then slowly declines to the low in May. This trend is constant across ten years of data. It is natural to have 1.85 to 2.99 Class B mishaps per 100,000 TFH occur per month.

The fact that the seasonal mishap rates are mostly within the “Band of Normal” reinforces that mishaps are settling to their natural rates of occurrence.

The AMIP is a tool designed to be utilized by AFSEC and USAF safety personnel at echelon to perform analysis and predictive modeling in accordance with AFI 91-202, Paragraph 5.4, requirements. Written in Python and capable of deployment on cloud environments, AMIP was designed to utilize standard data pulls from AFSAS and AFSEC to forecast the next 12 months of mishap incidence. AMIP can also perform analysis on a variety of aviation mishap phenomenon, such as engine mishaps or mission design series (MDS)-specific mishaps, by pulling data from AFSAS for that specific phenomenon. AMIP seamlessly integrates into, assists, and informs the Mishap Prevention Analysis processes specified in AFI 91-202, Paragraph 5.5.

References

International Civil Aviation Organization (2022) ICAO methodology for accident rate calculation and trending. SKYbrary. Accessed March 24, 2022, <https://skybrary.aero/articles/icao-methodology-accident-rate-calculation-and-trending>.

Kreisher O (2006) Toward zero mishaps. Air Force Magazine Accessed March 20, 2022, <https://www.airforcemag.com/article/1206mishaps/>.

United States Air Force Safety Center (2020) The U.S. Air Force mishap prevention program. Technical Report Air Force Instruction 91-202, Washington, DC, https://static.e-publishing.af.mil/production/1/af_se/publication/afi91-202/afi91-202.pdf.

Acknowledgments

Research occurring over such a long period and being as in-depth as this would not have been possible without the efforts of numerous individuals.

First and foremost, this research would not have been possible without the dedicated professionals of the U.S. Air Force Safety Center (AFSEC) and their efforts across the last year. I must thank Col Larry Nixon, USAF, Deputy Chief of Safety, for believing in the potential of this research enough to present it to the Air Force Chief of Safety for approval; for his relentless efforts to fulfill multiple, complex data requests; and his overall willingness to assist the research however possible. Without enthusiastic participation and support from Col Nixon, this research would never have been possible. I also owe gratitude to Lt Col Philip Bryant, USAF, HQ USAF Aviation Safety Issues, who picked up the mantle after the Initial Findings brief to be the primary point of contact with the Safety Center; without him, the Final Findings of this research would not have been possible. I thank the previous Air Force Chief of Safety, Maj Gen John Rauch, USAF, and the current Air Force Chief of Safety, Maj Gen Jeannie Leavitt, USAF, for the opportunity to conduct this research on behalf of AFSEC and for their support.

Next, I must thank those of the Naval Postgraduate School (NPS) who contributed to this research. To my Research Advisor, Professor Rudy Yoshida: thank you for taking on my research topic when many other professors were not interested. While this is not necessarily the most interesting topic, the mishaps that have occurred while this research was being conducted reinforced its unquestionable importance. To LTC Ross Schuchard, USA, Assistant Professor and fellow Operations Research/Systems Analyst (ORSA): thank you for your constant assistance as I developed the Aviation Mishap Incidence Predictor (AMIP) and its deployment options for AFSEC. You are the ORSA I hope to become. To my best friend here at NPS, CPT Tate Husemann, USA, and all the group partners across the duration of this curriculum: thank you for putting up with my hijacking of several group projects to conduct Proofs of Concept 1 and 2. To all those who proofread this thesis, the feedback of your red pens has developed me to communicate more effectively as an ORSA.

Finally, my entire family serves in the Department of Defense; their love and support have

been indispensable. To my Dad, a Department of the Air Force Civilian Chief of Safety whose efforts to develop an analysis tool in Excel made me think “there has to be a better way”: thank you for the countless hours of help and insight that made this research become far more effective than I could have made it alone. As ORSAs, we are taught that subject matter expertise can make or break your research. There is no doubt that your assistance has made this research effective enough to fulfill the role desired for it. To my Mom, a Department of the Air Force Civilian Family Advocacy Officer, and my step-mother, who served as a Key Spouse: thank you for all the emotional support you provided during the course of this research. I could not have sustained the effort and drive without it. To my brothers Drew and John: the fact that we serve has been a driving factor behind my desire to be an ORSA and improve the military however I can. You and the subordinates you lead deserve the best military possible. I love you all.

For all Safety Professionals who work tirelessly to preserve the lives of Service Members. It is my sincerest hope that this research assists you in your important mission.

CHAPTER 1:

Introduction

1.1 Project Origins and Problem Specification

Safety mishaps normally kill more U.S. service members annually than combat operations. In 2018, more service members died in aircraft mishaps than combat fatalities sustained in Afghanistan (Office 2019, p. 1). Commanders at all levels are ultimately responsible for the safety of their subordinates. The unsung heroes seeking to reduce the losses from safety mishaps are the commander's safety professionals. Working directly for commanders, the chiefs of safety and their staffs develop, oversee the implementation of, and evaluate the effectiveness of mishap prevention programs within their commands (United States Air Force Safety Center 2020, p. 21). Performing a dizzying myriad of duties, dedicated military safety professionals strive tirelessly to assist and advise commanders on how to best manage risk to reduce the likeliness of safety mishaps and prevent needless loss of life and equipment.

However, military safety programs are currently restricted to the rudimentary analysis of safety mishap data to identify emerging trends and incidence rate. This analysis, often conducted manually or utilizing unsuitable software solutions, fails to identify emerging trends in a timely manner and lacks the ability to accurately forecast the rate and trend of future aviation mishaps based on historical data. For example, U.S. Air Force (USAF) safety staffs often resort to building in-house analytical solutions in Microsoft Excel as a result of the lack of more sophisticated data analysis software and processes. This rudimentary and untimely analysis results in commanders setting arbitrary goals for decreasing the rate of mishaps for the next fiscal year (FY) or calendar year (CY). Using Air Force Safety Automated System (AFSAS) mishap and total flying hour data provided by the U.S. Air Force Safety Center (AFSEC), we construct time series analysis (TSA) Seasonal Decomposition and Auto Regressive Integrated Moving Average (ARIMA) predictive models that provide monthly trend, seasonal, and natural rates of occurrence for Class A and B military aviation mishaps and forecast the by-month incidence rate for the following year. These models will enable military safety professionals at echelon to focus their efforts on developing plans that

effectively target and counter emerging safety trends instead of expending precious man hours serving as data analysts.

1.2 Moral Imperative

According to the National Commission on Military Aviation Safety (2020, Front Cover), military aviation safety mishaps claimed 224 lives, 186 aircraft, and \$11.6 billion in damages between 2013 and 2020. To properly safeguard the lives and health of service members and government property, commanders and chiefs of safety require timely trend identification and incidence forecasting provided by sound data analytic processes to effectively target and combat safety trends.



Figure 1.1. Military Aviation Mishap Losses between 2013 and 2018. Commanders and chiefs of safety require timely trend identification and incidence forecasting provided by sound data analytic processes to effectively target and combat safety trends, protect lives, and preserve government property. Source: National Commission on Military Aviation Safety (2020, Front Cover).

1.3 Ties to USAF Mishap Analysis Program

This research will directly assist commanders and chiefs of safety in the performance of their duties. The USAF Mishap Prevention Program requires commanders and chiefs of safety to know the types of mishaps and mishap rates for their command, and directs them to develop risk mitigation strategies based on sound mishap analysis (United States Air Force Safety Center 2020, p. 77). U.S. Air Force instruction (AFI) 91-202, Paragraph 5.4, specifies that this analysis must:

- develop specific actions to reverse adverse trends;
- identify successes, problem areas, and trends to measure safety program effectiveness and guide prevention actions; and
- be conducted annually (United States Air Force Safety Center 2020, p. 77).

The models and tool developed by this research are designed to directly assist in the development of these annual safety programs by forecasting the next FY or CY of mishap incidence based on historical data. This research also provides commanders and chiefs of safety with indicators for when trends are exceeding normal rates of occurrence, as well as provides measures of performance (MOP) to measure safety program effectiveness. The model use cases detailed in Section 3.5.2 details how commanders and chiefs of safety can use this research within this paradigm. Sections 3.7.11 and 3.7.10 detail how to utilize this research to perform Mishap Prevention Analysis as defined by AFI 91-202, Paragraph 5.5.

1.4 Research Questions

Our research sought to answer the following questions:

1. Is it possible to predict the number of military aviation Class A and B mishaps using historical data from existing safety data systems?
2. Which regression methodologies are best suited to construct predictive models for military aviation safety mishaps?
3. What are the leading causes for Class A and B mishaps and are they changing across time?
4. Which aircraft types or models have the highest rates of Class A and B mishaps?

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2: Background

In this chapter, we begin with how aviation mishaps are defined and classified by the Department of Defense (DOD) because understanding these definitions are key to framing our research. Next, we examine previous research related to our topic, and then define the DOD Human Factor Analysis and Classification System (HFACS), the concept of Safety Privilege, and how both applied to our research. We conclude with an overview of the structure of the data provided from the Service safety systems of record and the data used for the research proofs of concept (POCs).

2.1 Key Mishap Definitions

Department of Defense instruction (DODI) 6055.07, dated 06 June 2011 with Change 1 dated 01 August 2018, defines a myriad of military aviation safety standards mishap notification, investigation, reporting, and record keeping. Enclosure 8, Figure 1, of the DODI (Figure 2.1) is a decision algorithm that confirms whether or not a military aviation mishap has occurred, and thereby provides a formal definition for military aviation mishaps (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics 2018, p. 34).

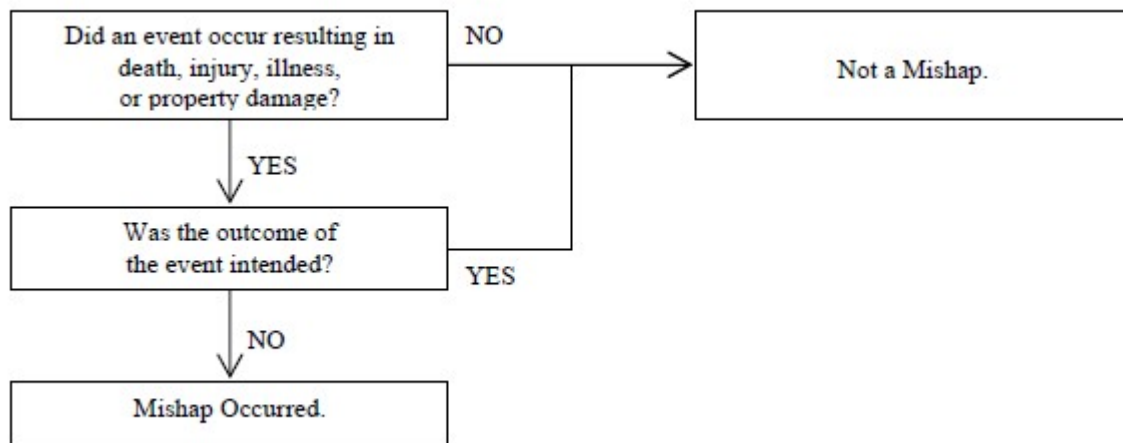


Figure 2.1. Decision Algorithm Defining Whether a Mishap has Occurred.
Source: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (2018, p. 34), Enclosure 8, Figure 1.

If mishap investigators answer “yes” to the first question and “no” to the second question, a mishap has occurred:

- Did an event occur resulting in death, injury, illness, or property damage?
- Was the outcome of the event intended? (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics 2018, p. 34)

Enclosure 8, Figure 2, of the DODI (Figure 2.2) defines the classes of aviation mishaps (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics 2018, p. 36).

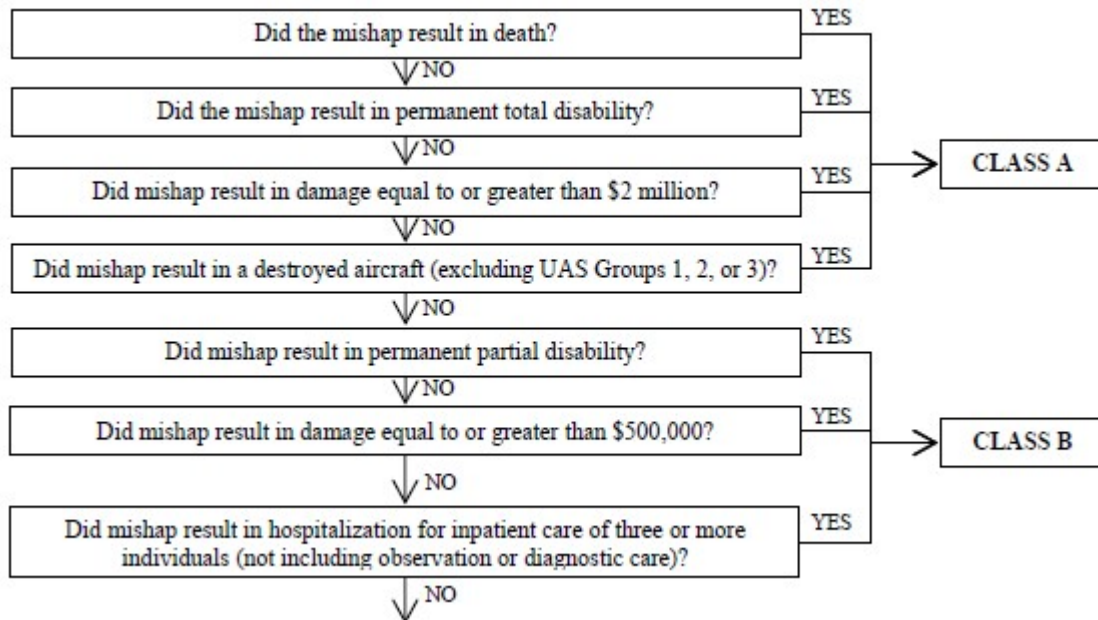


Figure 2.2. Decision Algorithm Defining Whether a Class A or Class B Mishap has Occurred. Source: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (2018, p. 36), Enclosure 8, Figure 2.

If mishap investigators answer “yes” to any of the following questions, the mishap is classified as a Class A mishap:

- Did the mishap result in death?
- Did the mishap result in permanent total disability?
- Did the mishap result in damage equal to or greater than \$2 million?
- Did the mishap result in a destroyed aircraft (excluding Unmanned Aircraft)? (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics 2018, p. 36)

If mishap investigators answer “yes” to any of the following questions after answering “no” to the above Class A mishap questions, the mishap is classified as a Class B mishap:

- Did the mishap result in permanent partial disability?
- Did the mishap result in damage equal to or greater than \$500,000?
- Did the mishap result in hospitalization for inpatient care of three or more individuals (not including observation or diagnostic care)? (Office of the Under Secretary of

Defense for Acquisition, Technology, and Logistics 2018, p. 36)

If mishap investigators answer “no” to the above questions, then the mishap may be classified as a lower class of mishap. As one can see from these definitions, Class A and B mishaps result in substantial losses for the USAF and DOD.

2.2 Literature Review

The literature on predicting the incidence of military aviation safety mishaps is relatively sparse. The sparsity of available research is indicative of two things: the difficulty of the research task and the novelty of the research area. For the review of available literature on the subject, we first examine the National Commission on Military Aviation Safety’s December 2020 Report to the President and the Congress of the United States, which not only communicates the importance of this research, but also highlights specific requirements for our research within safety programs across the Department of Defense. Then we examine related research conducted in the prediction of aviation mishap rates. The majority of recent research in the field has been conducted by Chinese researchers. The first research we examine was performed by the Chinese Air Force Engineering University’s Equipment Management and unmanned aerial vehicle (UAV) Engineering College. This research attempts to conduct stochastic forecasting of aviation safety mishaps using neural networks. While this research reveals an illuminating methodology, we explain why we feel neural networks are not suitable for our research and its application to U.S. military aviation mishaps.

2.2.1 The 2020 National Commission on Military Aviation Safety

The 2020 National Commission on Military Aviation Safety (NCMAS) report, submitted in December 2020 to the president and Congress, highlights the importance of our research. According to the Commission, during the conduct of its study from 2013 to 2020, over 6,000 military aviation safety mishaps occurred during training or routine operations (National Commission on Military Aviation Safety 2020, p. i). These mishaps cost the DOD and the Nation 224 lives, 186 aircraft, and \$11.6 billion in losses or damages (National Commission on Military Aviation Safety 2020, Front Cover). The costs of these largely preventable incidents are staggering. The fact that the DOD sustains such losses during training and

routine operations not only highlights the routinely dangerous nature of military service, but indicates that military aviation safety practices require an overhaul. Effective prediction of aviation mishap occurrence not only assists the development of predictive aviation safety programs but informs decision making processes that allocate resources and training, at both service and congressional levels.

NCMAS indicated the utilization of aviation safety data as one of the areas requiring improvement in military aviation programs. The Commission found that the “collection, processing, and analysis [of aviation safety data] must be improved before the data can be used effectively to improve military aviation safety across the DOD” (National Commission on Military Aviation Safety 2020, p. 15). Specifically, the Commission found that “the Department and Services have failed to take full advantage of commonly available trend analysis models and tools that could give them the ability to effectively collect, analyze, and operationalize safety data,” noting that “good data [and regular analysis] is the starting point for the identification of best practices and for making actionable military aviation safety decisions” (National Commission on Military Aviation Safety 2020, p. 15). The Commission recognized that, while the safety centers of each service have initiatives underway to improve their utilization of safety data, they “will require broader Service data collection, access, and analytical support to develop... effective predictive and actionable aviation mishap prevention program[s]” (National Commission on Military Aviation Safety 2020, p. 20). We hope that our research will provide the Air Force a foundation to perform advanced trend analysis and forecasting methodologies for aviation mishap incidence utilizing existing aviation safety data systems to better inform safety professionals and military leaders across the force.

2.2.2 Neural Network-Based Probability Forecasting of Aviation Safety Mishaps

Research jointly conducted by the Equipment Management and UAV Engineering College of the Chinese Air Force Engineering University and the Zhengzhou Regional Military Representative Office of Air Force examined the application of neural networks-based probability forecasting models for use in forecasting aviation mishaps. Zhu et al. (2021, p. 2) argue that “time series forecasting is only aimed at the time series of aviation [mishaps]” and that time series models ignore “the influence of correlation on [mishaps]” as well as

the impacts of uncertainty induced by causal factors, forecasting errors, and other forms of uncertainty. They further argue that probability forecasting with neural networks, where the initial weights and node output thresholds are generated randomly each time the neural network is trained, provides better inherent uncertainty in the model's predictions that better mirror the real-world uncertainty inherent in the occurrence of aviation mishaps (Zhu et al. 2021, p. 2).

Zhu et al. (2021, p. 4) apply a back propagation neural network (BPNN) in a two-part process to forecast yearly aviation mishap rates. According to Li et al. (2012, p. 553–554), a BPNN is “a multi-layer feedforward network trained according to error back propagation algorithms.” BPNNs utilize learning rules that “adopt the steepest descent method in which the back propagation is used to regulate the weight and threshold values of the network to achieve the minimum error” (Li et al. 2012, p. 553). The first part of the BPNN generates point value forecasts for the yearly mishap rate by generating 10 most likely forecast values (Zhu et al. 2021, p. 4). The second part of the BPNN fits a probability density curve to the point forecasts and their corresponding error values, and generates an interval estimate by integrating across the constructed probability curves to produce the upper and lower confidence bounds on a 90% confidence level (Zhu et al. 2021, p. 5). This procedure is depicted in Figure 2 of their research (Figure 2.3).

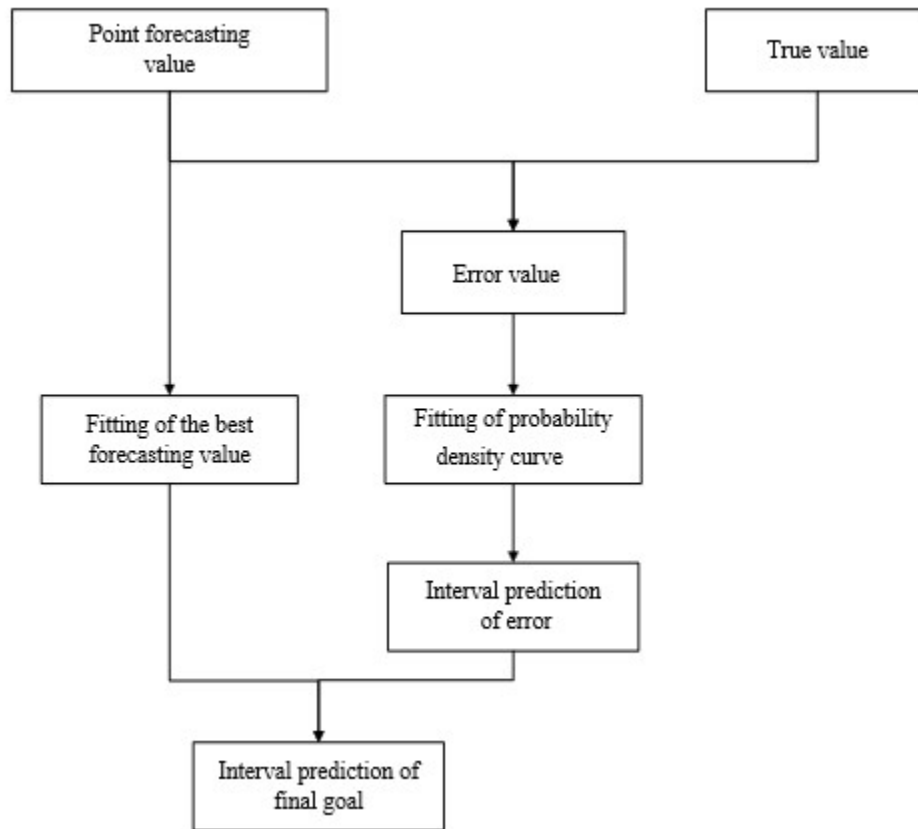


Figure 2.3. Flow Chart Depicting Error Back Propagation for Forecasting Aviation Mishap Rates. Source: Zhu et al. (2021, p. 6), Figure 2.

While the use of BPNN may improve forecasting accuracy, reduce forecasting error, and attempt to better capture the uncertainty inherent in mishap incidence, we feel the disadvantages inherent to neural networks negates their desirability for application to DOD aviation mishap decision making processes. The numerous, complex interactions occurring within a neural network severely decrease interpretability of any results produced and removes the ability of analysts to validate the calculations occurring within the network. Often, neural networks will randomly and inexplicably drop weight values or remove interactions as the network attempts to minimize its loss function values, producing erroneous results which are difficult to troubleshoot. Additionally, calibrating neural networks to training data is incredibly difficult for subject matter experts, and these calibrations change according to the data for which the network is being trained. Such a tool would be incredibly difficult

to implement within the DOD, where such subject matter expertise is not available at the Service safety center level, and where interpretability of the results by both analysts and decision makers is essential to make actionable aviation safety plans.

2.3 Human Factor Analysis and Classification System

The DOD has developed a classification system of mishap causes that is extremely useful for the analysis of aviation mishaps. According to the United States Air Force Safety Center (2019, p. 1), the HFACS model “presents a systematic, multidimensional approach to error analysis and mishap prevention and that mishap investigators must use the system in mishap analysis”. They are the DOD’s implementation of James Reason’s “Swiss Cheese” model, a systematic approach that describes “four levels within which active failures and latent failures may occur during complex operations” (United States Air Force Safety Center 2019, p. 1). Active failures are defined as “the actions or inactions of individuals that are believed to cause or contribute to a mishap” (United States Air Force Safety Center 2019, p. 1). Latent failures are defined as “pre-existing conditions within an organization which indirectly affect the sequence of mishap events” (United States Air Force Safety Center 2019, p. 1). Human factors are further categorized as being either mishap-level or person-level. This model is depicted in Figure 1 of United States Air Force Safety Center (2019, p. 2), shown in Figure 2.4.

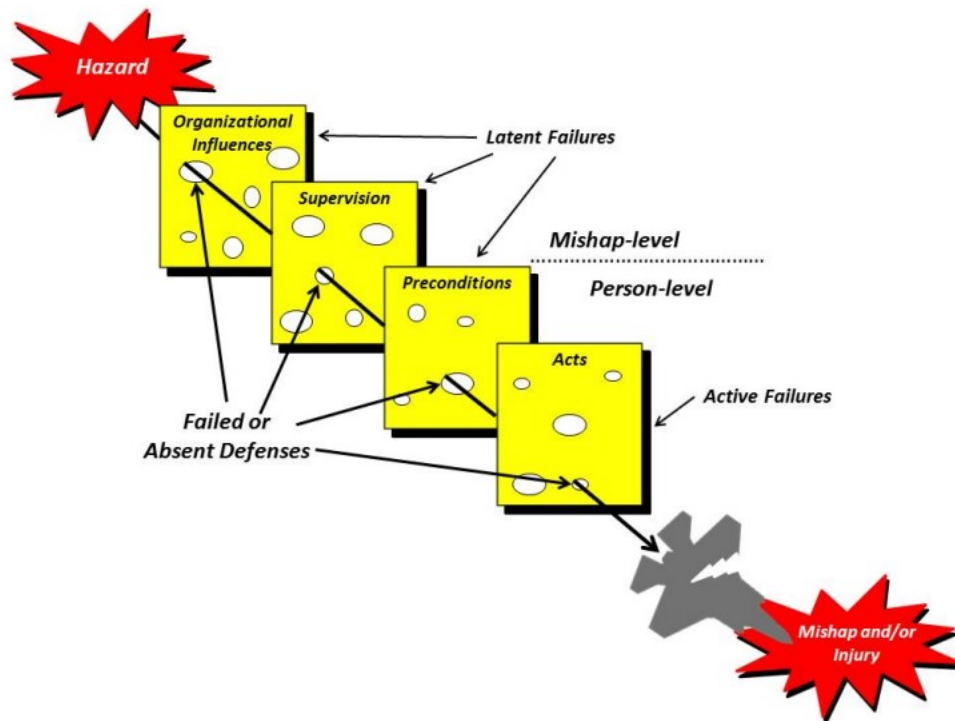


Figure 2.4. Swiss Cheese Model for Aviation Mishaps. Source: United States Air Force Safety Center (2019, p. 2).

The HFACS assign a unique five-character alphanumeric code to each of the categories of human factors which could either cause or contribute to a mishap under any of the Swiss Cheese model's four levels, resulting in 108 distinct codes, as in Figure 2.5. This system can be extremely useful for classification problems, which are one of the two typical Operations Research modeling problems. This system can also be harnessed to link potential causes with mishap forecasts, providing human factors for DOD commanders and safety professionals to target in yearly safety programs.

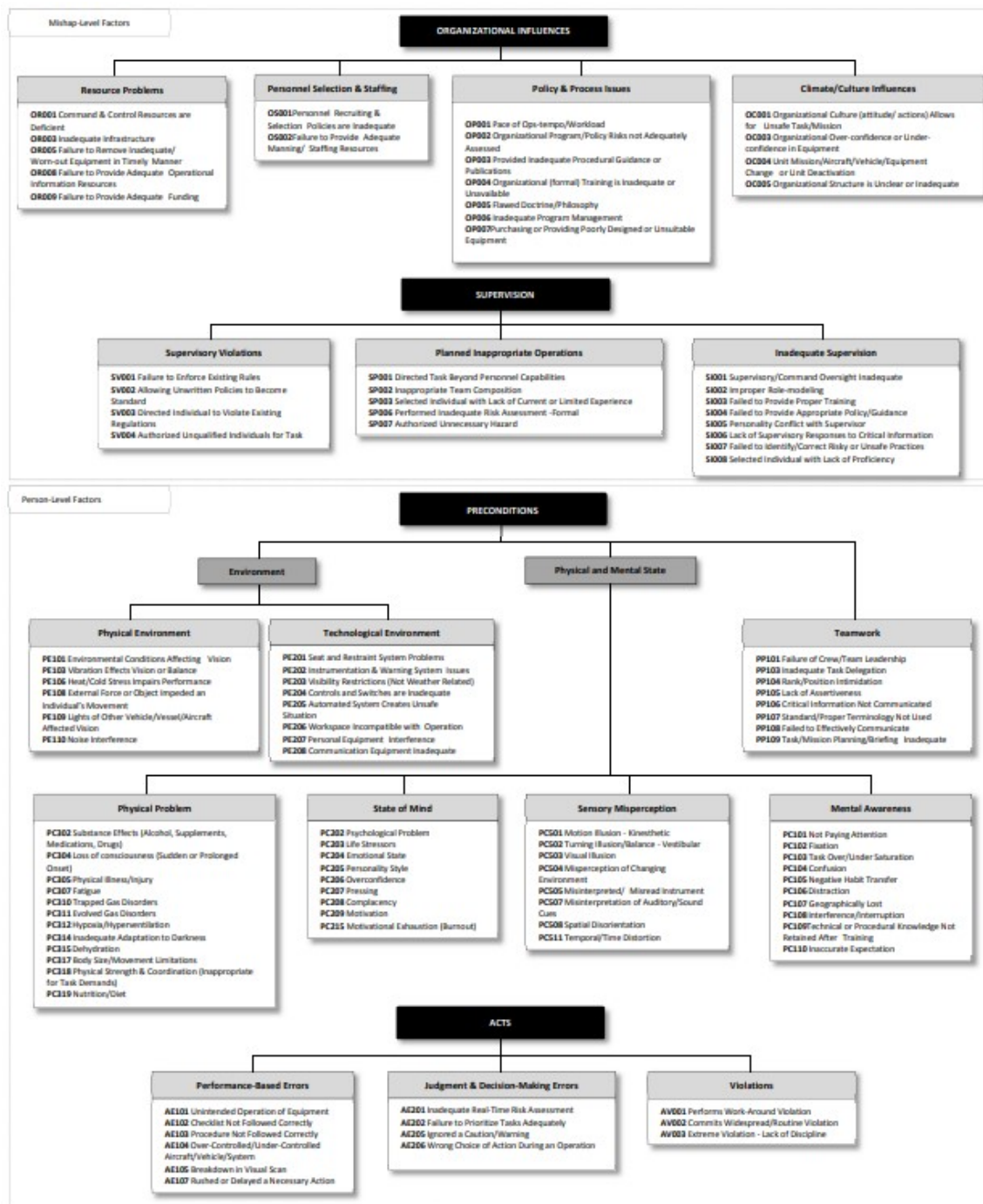


Figure 2.5. HFACS Code Tree. Source: United States Air Force Safety Center (2019, p. 22).

2.4 Privileged Safety Information

privileged safety information (PSI) played a key role in the scope of research. AFI 91-204 defines PSI as several categories of information, including “analysis, findings, conclusions, causes, recommendations, other findings and recommendations of significance” of safety investigators (United States Air Force Safety Center 2021b, p. 31). The USAF treats this information as limited use and limited access “based on a national defense need for rapid and accurate assessment of the causes of mishaps to prevent a recurrence and maintain mission readiness” (United States Air Force Safety Center 2021b, p. 31). Since the idea of PSI is rooted in U.S. case law, the release and use of PSI is closely controlled.

This close control, under the requirement to protect PSI coupled with the unclassified nature of this research, limited the analysis that could be conducted by the research team. In fact, the objective of the third research question was completely unobtainable after the separation of the HFACS data from the individual mishap data. To safeguard the PSI in this unclassified research context, the causes of each mishap not only had to be completely disconnected from individual mishaps, but the format of the HFACS data had to be such that the causes could not be reconnected to their individual mishaps. Therefore, the HFACS data included no date information (see Section 2.5). The lack of this information prevented the development of TSA models to forecast time-associated mishap causes. It also prevented seasonal decomposition analysis of HFACS occurrences to determine if there are seasonal trends to mishap causes. While this research is completely feasible given the current composition and quality of USAF mishap data, such research must be conducted with PSI release restrictions under PSI safeguarded environments. It was determined that any analysis performed on the HFACS data would not add to the analysis already being performed by USAF analysts.

2.5 United States Air Force Safety Automated System Data

The AFSEC provided three data sets at the request of the research team: individual mishap data; mishap HFACS data; and the number of active pilots (United States Air Force Safety Center 2021a). The individual Mishap and HFACS data sets were pulled directly from the AFSAS, the USAF system of record for all mishap data. The number of active pilots was provided by the U.S. Air Force Personnel Center (AFPC). Due to the importance of Class A and B mishaps, AFSAS data for these classes is reviewed by AFSEC for completeness and integrity. Given the quality of the data for these classes and their importance to the USAF,

research analysis was restricted to Class A and B mishaps. Complete data for FY 2010 to FY 2020 was provided, as well as all data available for FY 2021 at the time of the request. Analysis was therefore conducted on more than ten years of data ranging from 01 October 2010 to 31 March 2021. The duration of these data sets must be emphasized, as any mishap trend analysis conducted for ten years of data is substantial, despite the low number of mishaps per month. Analysis was conducted on all Aircraft Flight, Aircraft Flight-Related, and Aircraft Ground Class A and B mishaps, regardless of cause, and included both manned and unmanned aircraft.

The Mishap data set contained 832 total mishaps: 310 Class A mishaps and 522 Class B mishaps. It was drawn directly from AFSAS using a tailored query and contained the following columns:

- Mishap Class
- Date
- MDS Type
- MDS Aircraft
- Category
- Mishap “One-Liner” Summary

Mishap Class is a categorical variable with values of either “A” or “B.” mission design series (MDS) is defined by AFI 16-401 as “the official designation for aerospace vehicles” used to represent “a specific category of aerospace vehicles for operations, support, and documentation purposes” (United States Air Force 2020, p. 6). MDS Type is a categorical variable with values of the kind of primary mission the aircraft performs (“Fighter/Bomber,” “Transport,” etc). MDS Aircraft is a categorical variable with the value of the model and variant of the aircraft (“A-10C,” “B-1B,” etc.). Category is a categorical variable with the values of what kind of mishap occurred. The mishap “One-Liner” is one line summary that explains what occurred during the mishap.

The HFACS data set was not utilized, as it had no time elements associated with each record, preventing the performance of trend analysis. Therefore, the structure of the data set will not be discussed in detail.

The Number of Active Pilots data set was requested by the research team during proof of

concept (POC) 1 while it was investigating simple linear regression (SLR) as a possible modeling method. It was felt by the research team that the number of active pilots each year would be beneficial to add as a predictor in an SLR model. However, as research pivoted toward TSA, this data set was no longer required. Therefore, the structure of the data set will not be discussed in detail.

2.6 National Transportation Safety Board and Federal Aviation Administration Data for Proofs of Concept

A data set in similar format to the AFSAS data was required by the research team to conduct the POCs while AFSEC collected the requested data. Open-source National Transportation Safety Board (NTSB) individual mishap data was sourced through the data website Kaggle (Samaha 2021). The entire data set contains nearly 86,000 mishaps from 1948 to present day, with the following columns that were completed as applicable to the mishap:

- Accident ID Number
- Event Date
- Injury Severity
- Aircraft Damage
- Aircraft Make
- Aircraft Model
- Federal Aviation Regulation (FAR) Description (Category of Operation)
- Purpose of Flight
- Air Carrier
- Total Fatal Injuries
- Total Serious Injuries
- Total Minor Injuries
- Total Uninjured
- Weather Condition (“VMC” or “IMC”)
- Broad Phase of Flight
- Accident Class

The contents and format of the NTSB data is very similar to that of the AFSAS data. The NTSB data is also more detailed in many ways. However, the NTSB does not classify the

severity of a mishap in the same manner as the USAF. Therefore, the research team had to construct the “Accident Class” field to the data according to the following definitions:

- Class A: Any incident resulting in the complete destruction of the aircraft or a fatality.
- Class B: Any incident resulting in damage to the aircraft or injury to an individual.

These definitions were developed by the research team to roughly apply the delineations of the DODI 6055.07 Class A and B mishap definitions to the form of data present in the NTSB mishap data.

Total Flight Numbers data was sourced from the Federal Aviation Administration (FAA) OPSNET database (Federal Aviation Administration 2021). The database was queried to provide the total number of flight operations by year by FAA region from 01 January 2000 to 31 December 2020. This resulted in a data set of nearly 500 lines. Due to OPSNET query limitations, the data was pulled in two 10-calendar year blocks: 2000–2010 and 2011–2020. The data contained the following columns:

- Calendar Year
- DDSO Service Area
- Class
- IFR Air Carrier Operations
- IFR Air Taxi Operations
- IFR General Aviation Operations
- IFR Military Operations
- IFR Total Operations
- VFR Air Carrier Operations
- VFR Air Taxi Operations
- VFR General Aviation Operations
- VFR Military Operations
- VFR Total Operations
- Total Operations

In an effort to maximize data quality, analysis was confined to the time period ranging from 01 January 2000 to 31 December 2020, and only examined FAR Part 121: Commercial Aviation and FAR Part 91: General Aviation mishaps. This reduced data contained 20,347

mishaps: 1,662 Class A mishaps and 18,686 Class B mishaps. The combined data set resulted in the following columns:

- Year
- Total Number of instrument flight rules (IFR) Flights
- Total Number of visual flight rules (VFR) Flights
- Total Number of Flights (Operations)
- Total Number of Class A Mishaps
- Total Number of Class B Mishaps

A deeper exploratory data analysis (EDA) will be discussed in Chapter 3 while discussing POC1.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

3.1 Overview of Predictive Modeling Methodologies

This overview is designed to give the reader a basic understanding of predictive modeling methodologies, which is confined to only those used in the research. More importantly, this overview attempts to provide the reader with an understanding of how the predictive performance of these models is evaluated. This knowledge is crucial for understanding our research.

Predictive modeling methodologies attempt to estimate the relationship of a variable of interest (called as a response variable or target variable) Y to one or more X_i input predictor variables. The “variable of interest” is normally some outcome or resulting value of a phenomenon that we wish to analyze or to model given the values of input factors. For example, in this research, the “value of interest” is the monthly number of mishaps or the mishap rate per total flying hours, given the input mishap factors. The mathematical complexity of this estimated relationship varies depending on the model used. This complexity ranges from the straightforward SLR model to complex ensemble models like ARIMA.

Prior to fitting any models, the data must be split into two or three subsets: Training, Validation, and Testing. At a minimum, the data must be split into Training and Testing subsets, with the Validation subset being optional but considered a best practice. Each subset has a specific purpose. The Training subset is the part of the data which the models are fit against in order to estimate the parameters for a particular phenomenon of interest. The Training subset can contain as much as 80% of the original data set, depending on the principle used to calculate its size (Wijaya 2020). The Testing data is used to evaluate forecasting performance after the model is fully trained, and normally consists of the remainder of the original data whenever a Validation set is not used. The optional Validation data is used in certain modeling methodologies to evaluate (and in some cases inform) the fitting of the model as it is being trained. The use of the Validation data preserves the integrity of the Testing data for evaluation. The Testing data should only be used to evaluate

model forecasting performance after analysts believe it is trained appropriately. The size of the subsets is selected by the analyst, and depends on the quality of the data and the amount of data required to adequately train the model for a particular phenomenon.

The method for splitting the data changes depending on the modeling methodology. For SLR and most other regression methodologies, the data is split randomly into the Training, Testing, and Validation subsets (Hyndman and Athanasopoulos 2018, ch. 3.4). For example, if you want a 75–25 Training-Testing split of your data, 75% of the observations are randomly indexed into the Training subset, with the remaining 25% of the observations indexed into the Testing subset. However, in TSA, the time-series data must remain as continuous blocks of time comprised of whole periods when split (Hyndman and Athanasopoulos 2018, ch. 1). The period is the block of time the TSA model is attempting to forecast. For example, say you have six years of time-series data consisting of 72 monthly observations and wish to construct a TSA model for a phenomenon with a 12-month period. The first four years of the data can become the Training subset with the subsequent two years becoming the Testing subset. In this way, any trend or seasonality characteristics inherent to the period are captured by the model.

3.1.1 The Quick Basics of Linear Regression

SLR is the cornerstone of regression models for predictive modeling methodologies. In its most basic form, SLR models assume some form of linear relationship exists between the variable of interest (response variable) Y and the X_i predictor variables (Hyndman and Athanasopoulos 2018, ch. 5.1):

$$Y_t = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon_t$$

where:

- $Y_t \in \mathbb{R}$ = the variable of interest at time t for $t = 0, \dots, T$ time steps
- $X_i \in \mathbb{R}$ for $i = 1, \dots, n$ = the i th predictor variable for input category i of n input categories
- $\beta_0 \in \mathbb{R}$ = intercept for Y_i when all $X_i = 0$
- $\beta_i \in \mathbb{R}$ = intercept for the corresponding X_i ; represents the average predicted change

- in Y_i resulting from a one unit increase in the corresponding X_i
- $\epsilon_t \in \mathbb{R}$ = the amount of random error present in Y_t .

More complex SLR models assume a more complex relationship exists between the X_i predictor variables and the Y variable of interest. These more complex relationships can take the form of:

- nonlinear values, such as X_i^2 or X_i^3 ; or
- interactions between X_i predictors, such as $(X_1 \times X_2)$ or $(X_1 + X_2)$.

When training (also known as “fitting”) an SLR model, the β_n coefficients are estimated by minimizing the root mean squared error (RMSE) value for the relationship. We discuss RMSE further in Section 3.1.2.

The proper estimation of SLR model parameters hinges on meeting the requirements of five key modeling assumptions. These assumptions must be satisfied for the model to be valid:

- the linear or nonlinear relationship estimated by the model is a reasonable approximation of reality;
- the errors have a mean of zero;
- the errors are not correlated;
- the errors are not related to the predictor variables; and
- the errors are Normally distributed with a constant variance (Hyndman and Athanasopoulos 2018, ch. 5.1).

These modeling assumptions played a critical role in the models developed in POC1.

3.1.2 Analyzing Linear Regression Model Performance

RMSE was the forecast error measure utilized to evaluate model performance in POC1. The closer the error values are to zero for a given SLR model, the more accurate the model (Hyndman and Athanasopoulos 2018, ch. 3.4).

Testing an SLR model will result in a series of estimated $\hat{Y}_t$ outcomes for $t = 0, \dots, T$ time steps. Note the use of the hat symbol^ here to denote an estimated value. To measure the performance of the model, these estimated outcomes are checked against the actual Y_t

outcome variables from the Testing data. This measurement results in the forecast error, defined by Hyndman and Athanasopoulos (2018, ch. 3.4) as the difference between an observed value and its forecast. RMSE is one of the most commonly used measurements of forecast error (Hyndman and Athanasopoulos 2018, ch. 3.4):

$$RMSE = \sqrt{mean((Y_t - \hat{Y}_t)^2)}$$

where:

- $Y_t \in \mathbb{R}$ = the variable of interest Y at time step t in the Testing data
- $\hat{Y}_t \in \mathbb{R}$ = the forecasted value $\hat{Y}$ at time step t by the model.

3.1.3 The Quick Basics of Time Series Analysis

A significant quantity of military data is collected as and exists in time-series format, making the utilization of TSA an important consideration for military analysts. Time-series data is composed of observations collected sequentially over time that are normally observed at regular intervals (hourly, daily, weekly, monthly, quarterly, yearly, etc.) (Hyndman and Athanasopoulos 2018, ch. 1). While military aviation mishaps occur irregularly, their incidence is consolidated into monthly and annual counts and rates. The goal of TSA is to estimate how the sequence of observations will continue into the future (Hyndman and Athanasopoulos 2018, ch. 1). In our research, we wished to forecast the USAF monthly number of mishaps and monthly mishap rate for a future 12-month period given a number of years of historical mishap data from AFSAS.

There are a multitude of TSA models. Our research tested Seasonal Decomposition, Holt-Winters, and ARIMA ensemble models in POC2. We will briefly discuss how each model works and what each model has to offer with regard to TSA forecasting.

The Seasonal Decomposition model breaks time-series data into three components for forecasting: Trend, Seasonality, and Remainder. Forecasts are the result of either an Additive Model:

$$Y_t = S_t + T_t + R_t$$

or a Multiplicative Model:

$$Y_t = S_t \times T_t \times R_t$$

where:

- $\hat{Y}_t \in \mathbb{R}$ = the variable of interest at time t for $t = 0, \dots, T$ time steps
- $S_t \in \mathbb{R}$ = the Seasonality value at time t ; representative of the seasonal fluctuation within the period
- $T_t \in \mathbb{R}$ = the Trend value at time t ; representative of the increasing or decreasing rate of the phenomenon
- $R_t \in \mathbb{R}$ = the Remainder value at time t ; representative of the randomness inherent to a phenomenon outside of its Seasonality and Trend (Hyndman and Athanasopoulos 2018, ch. 6.1).

The Additive Seasonal Decomposition model is used with phenomenon where the magnitude of seasonal fluctuations or trend does not vary with the level of the time-series data (Hyndman and Athanasopoulos 2018, ch. 6.1). The Multiplicative Seasonal Decomposition model is used with phenomenon where the fluctuations appear to be proportional to the level of the time-series data (Hyndman and Athanasopoulos 2018, ch. 6.1). Section 3.7.7 discusses the use of each in the final product of our research. The Seasonal Decomposition model is useful when wishing to identify the peak and low months of occurrence for a given phenomenon, as well as the overall trend for the occurrence of a phenomenon.

The Holt-Winters model utilizes exponential smoothing of trend data and seasonality to make forecasts. It attempts to fit three parameters for three smoothing equations:

- smoothing parameter α for Level equation l_t ;
- smoothing parameter β for Trend equation b_t ;
- smoothing parameter γ for Seasonality equation s_t (Hyndman and Athanasopoulos 2018, ch. 7.3).

Each equation is essentially a weighted average of each component at time step t smoothed exponentially so that older observations have less weight than more recent observations. The Holt-Winters model also has Additive and Multiplicative variations. As an exponential smoothing model, Holt-Winters is useful for modeling phenomenon where the effects of older observations in the time-series do not have as much of an effect on the forecasted outcome as more recent observations.

The ARIMA model is an ensemble model that incorporates two models that aim to estimate the autocorrelations in time-series data: Auto-Regressive models and Moving Average models. Fitting of an ARIMA model attempts to estimate three parameters:

- p , the order parameter for the Auto-Regressive component;
- d , the degree of first differencing involved;
- q , the order parameter for the Moving Average component (Hyndman and Athanasopoulos 2018, ch. 8.5).

ARIMA models can therefore be identified as $ARIMA(p, d, q)$ models, with the p , d , and q values identifying how the ARIMA model is trained to a given phenomenon. ARIMA models can be configured to include or ignore seasonality of a phenomenon. According to Hyndman and Athanasopoulos (2018, ch. 8.5), fitting of these parameters is difficult, and therefore it is advised to use programming packages to estimate these values automatically. This research makes use of R's native `auto.arima()` function and Python's `auto.arima()` function from the `pmdarima` library to perform estimation of these parameters. The use of ensemble models like ARIMA allows for the development of more complex models that are more capable of forecasting complex phenomenon than simpler modeling methodologies.

Any of these models can forecast in a Naive form. The Naive form of these models simply predicts only using the $t - 1$ time step for the Y_t forecast without any additional information from other previous time steps. A Naive model can also be built, where the Y_t forecast value is simply the Y_{t-1} value.

3.1.4 Analyzing Time Series Model Performance

The closer the error values for a given TSA model are to zero, the more accurate the model. Our research utilized mean absolute percentage error (MAPE) and mean absolute

scaled error (MASE) metrics to perform the primary model performance evaluation in R. However, symmetric mean absolute percentage error (SMAPE) was used to perform secondary model performance evaluation in Python due to its built-in deployment in the `pmdarima` package used to fit the ARIMA models. Ultimately, MASE is the best measure of predictive performance for the TSA models developed for this research.

MAPE is a percentage measure of performance that is often used to compare forecast performance between different TSA models due to its independence from the scale of the data (Hyndman 2006, p. 45). The comparison of multiple, competing TSA models in this research necessitated the use of MAPE over the standard mean squared error (MSE) and RMSE measurements of forecast error normally used for model performance evaluation. As described by Hyndman (2006, p. 44), MSE and RMSE are not meaningful for assessing a method's accuracy across multiple series because its error measurements are on the same scale as the data. MAPE is calculated as:

$$\epsilon_{MAPE} = mean(|\frac{100(Y_t - F_t)}{Y_t}|)$$

where:

- $\epsilon_{MAPE} \in [0, 100]$ = the mean absolute percentage error for the TSA model being evaluated
- $Y_t \in (-\infty, \infty)$ = the actual outcome value at time step t
- $F_t \in (-\infty, \infty)$ = the forecasted value at time step t predicted by the TSA model (Hyndman 2006, p. 45).

However, MAPE has two disadvantages. The first disadvantage is that MAPE penalizes positive error values heavier than it does negative error values (Hyndman 2006, p. 45). This has led to the widespread use of SMAPE, a derivative of MAPE which corrects for this issue:

$$\epsilon_{SMAPE} = mean(\frac{200|Y_t - F_t|}{Y_t + F_t})$$

where:

- $\epsilon_{SMAPE} \in [-100, 100]$ = the symmetric mean absolute percentage error for the TSA model being evaluated
- $Y_t \in (-\infty, \infty)$ = the actual outcome value at time step t
- $F_t \in (-\infty, \infty)$ = the forecasted value at time step t predicted by the TSA model (Hyndman 2006, p. 45).

Second, MAPE and SMAPE have poor performance with data where zero values can appear in the time series. Zero values in the time series can skew the MAPE distribution, causing the MAPE values to be unusually large, and can cause MAPE values to be infinite or undefined (Hyndman 2006, p. 44). This disadvantage is especially relevant to this research, where it is possible (and actually desirable) for mishaps to not occur during any given month.

In contrast to MAPE, MASE is a scale-free measure of performance that can be used to compare forecast performance of TSA models both on a single time series and across multiple time series (Hyndman 2006, p. 46). MASE accomplishes its scale-free property by being scaled by the error of the naive forecast for the time series, as can be seen in the denominator of the MASE equation:

$$\epsilon_{MASE} = mean(|\frac{Y_t - F_t}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}|)$$

where:

- $\epsilon_{MASE} \in [0, \infty)$ = the mean absolute scaled error for the TSA model being evaluated
- $Y_t \in (-\infty, \infty)$ = the actual outcome value at time step t
- $F_t \in (-\infty, \infty)$ = the forecasted value at time step t predicted by the TSA model
- $Y_i \in (-\infty, \infty)$ = the i th actual outcome value in the data
- $Y_{i-1} \in (-\infty, \infty)$ = the $i - 1$ th actual outcome value in the data
- $n \in [0, \infty)$ = the total number of actual outcomes Y in the data (Hyndman 2006, p. 45).

3.2 Proof of Concept 1: Simple Linear Regression

As the cornerstone of regression analysis, SLR is the natural starting point for building predictive models. POC1, conducted during July 2021, set out to determine the suitability-

ity of applying SLR models for predicting DOD aviation mishaps utilizing the current configuration of AFSAS data.

POC1 sought to answer the following questions:

1. Is it possible to predict the number of Aviation Class A and B mishaps for a given year?
2. Which predictors provide the most reliable predictions for the number of Class A and B mishaps?

POC1 explored these research questions simultaneously through construction of two different models for Class A and B mishaps. POC1 utilized NTSB mishap data similar in format and structure to the AFSAS data combined with FAA total operations data. POC1 analysis and modeling was performed using R.

Key Definitions

As the NTSB data is not exactly the same as the AFSAS data, and in an effort to mold SLR methodologies to the structure of the data, the following definitions were applied to POC1 specifically:

- Class A Mishap: Any incident resulting in the complete destruction of the aircraft or a fatality.
- Class B Mishap: Any incident resulting in damage to the aircraft or injury to an individual.
- Visual Flight Rules (VFR): Flight conditions with line of sight visibility of 3 miles under 10,000 feet or 5 miles over 10,000 feet.
- Instrument Flight Rules (IFR): Flight conditions with line of sight visibility less than VFR requirements; includes flying in inclement weather conditions or at night.

3.2.1 Exploratory Data Analysis

The NTSB individual mishap data and FAA total operations data were both described in detail in Chapter 2. A Python script was written by the authors to load the mishap data into a Pandas dataframe, standardize categorical predictor entries, and classify each incident as either Class A or B according to the definitions stated earlier. The counts of Class A

and B mishaps by year were then calculated. The Python script also loaded the FAA total operations data into a separate Pandas dataframe and extracted the operation subtotals for each year. This data was then merged with yearly counts of Class A and B mishaps from the individual incident data. This resulted in the complete POC1 data set containing 21 observations with four possible predictors:

- Year;
- IFR.Total;
- VFR.Total; and
- Total.Operations.

The Pandas dataframe was then exported as a .csv file to be used separately in R for model development. The data from both the NTSB and FAA data sets had to be combined in this way due to the nature of SLR models and how they can only operate on numerical predictors. The reduction of the data to 21 observations was one of the first recognized shortcomings of using SLR models to predict mishap incidence.

EDA of the individual mishap data set reveals 1,662 Class A and 18,686 Class B mishaps occurring between 01 January 2000 and 31 December 2020. Of these, 953 occurred under IFR weather conditions, 19,298 occurred under VFR conditions, and 97 occurred under unknown weather conditions. 93% of incidents occurred with Part 97: General Aviation aircraft, which concurs with the large portion of aircraft being of the General Aviation type. The remaining 7% of incidents occurred with Air Carrier and Air Taxi/Commuter aircraft, and helicopters. Of note, 6 incidents occurred involving Armed Forces aircraft.

EDA of the occurrence of Class A mishaps (Figure 3.1) displays a bimodal distribution across the 20 years of data. Note how the trend line fails to connect all data points. This means fitting a model exactly to the Class A mishap data using SLR will be difficult. Of note, however, is the general upward trend in the occurrence of Class A mishaps, especially from 2013 to 2020. EDA of the occurrence of Class B mishaps (Figure 3.2) displays a sharp increase in the rate of mishaps from 2000 to 2012, then a steady decline from 2013 to 2020.

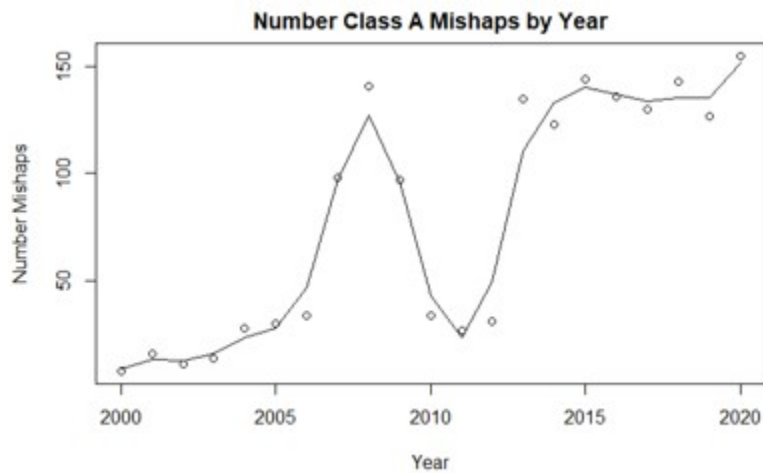


Figure 3.1. Occurrence of Class A Mishaps in NTSB Data. The failure of the trend line to pass through all points of the distribution foreshadows the difficulty of SLR models to fit this bimodal data.

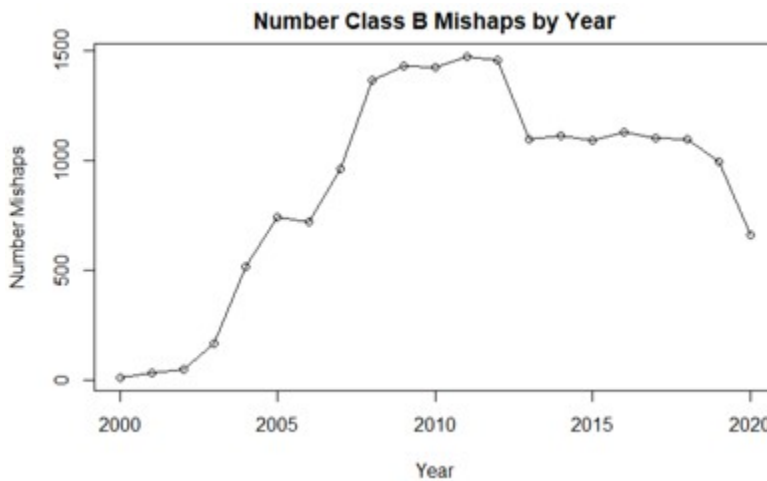


Figure 3.2. Occurrence of Class B Mishaps in NTSB Data

The FAA conducts between 42 million and 68 million flights (operations) per year (Figure 3.3). Approximately 66.5% of those operations are IFR, with the remainder being VFR. It is recognized by the authors that this data set does not entirely capture all the flights conducted in the United States, as many General Aviation flights occur without direct interaction of pilots with the FAA. However, this data set gave enough of a complete picture to fully

demonstrate the viability of the POC.

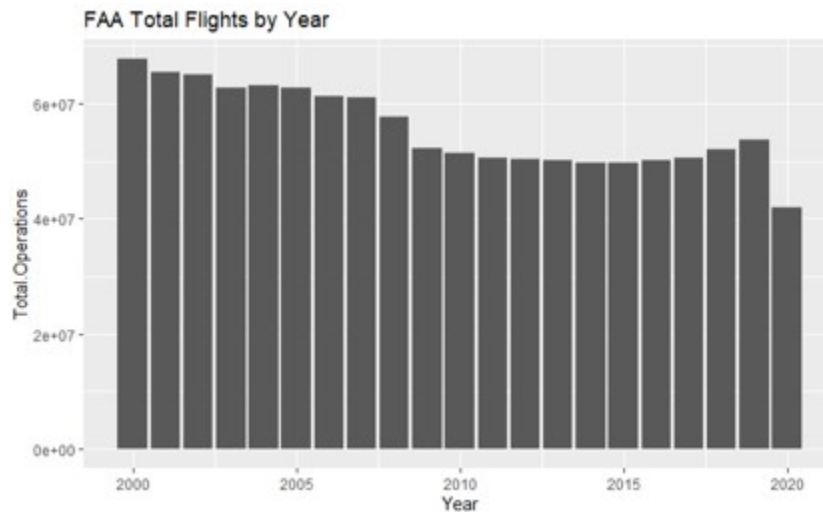


Figure 3.3. Histogram of FAA Total Operations by Year

Examination of the Pairs Plot for the aggregate data reveals a sparse data set that does not conform well to the three SLR modeling assumptions (Figure 3.4). The presence of multicollinearity between nearly all the predictors is of particular importance. The extent of this multicollinearity indicates SLR models will be particularly unsuited for predictive models using this data.

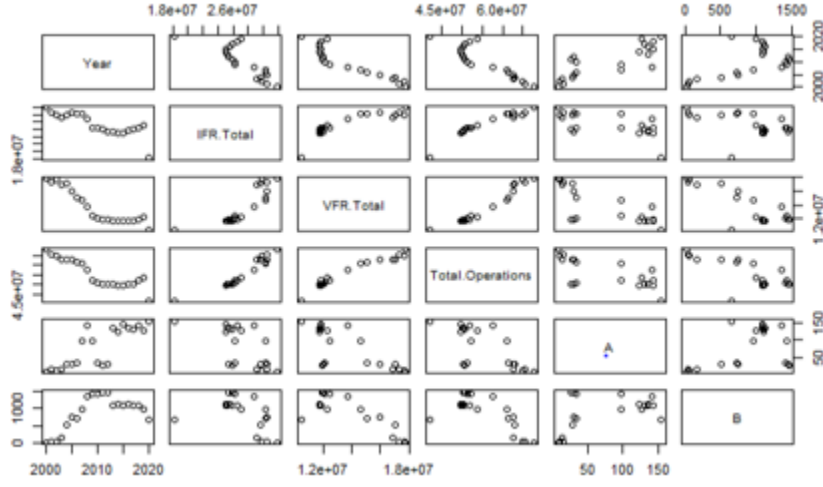


Figure 3.4. Pairs Plot of POC1 Data Indicating a Sparse Data Set with Multicollinearity

For development of the models, the POC1 data was split 75/25 to form training and testing data sets to produce both training and testing RMSE values as the primary means of model evaluation.

3.2.2 Predicting the Number of Mishaps — Initial SLR Models

All available predictors from the POC1 data set were fed into the initial models to establish a baseline for follow-on evolutions. The initial model for predicting Class A mishaps was formulated as follows:

$$Y_A = \beta_0 + \beta_1 X_{Year} + \beta_2 X_{Total.Operations} + \beta_3 X_{VFR.Total} + \beta_4 X_{IFR.Total}$$

where:

- $Y_A \in [0, \infty)$ = the number of Class A mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = Year coefficient
- $X_{Year} = \langle 2000, \dots, 2020 \rangle$ = vector of Year predictors

- $\beta_2 \in (-\infty, \infty)$ = Total Operations coefficient
- $X_{Total.Operations}$ = vector of the number of Total Operations predictors for each year
- $\beta_3 \in (-\infty, \infty)$ = VFR Operations coefficient
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year
- $\beta_4 \in (-\infty, \infty)$ = IRF Operations coefficient
- $X_{IRF.Total}$ = vector of the number of IRF Operations predictors for each year.

The performance results of the initial Class A model show a large difference between the Training and Testing RMSE, the first and second numbers depicted in Figure 3.5. Negative predictions for the number of Class A mishaps against the test data set's actual Y values in Figure 3.5 reveals a model that is poorly fit to the data. variance inflation factor (VIF) analysis depicted at the bottom of Figure 3.5 reveals high multicollinearity between Total.Operations, VFR.Total, and IRF.Total predictors. The initial Class A model does not conform well to the three modeling assumptions; transformations are required.

```
[1] "Training RMSE: 32.0947380473438"
[1] "Testing RMSE: 403.756707841355"
[1] "Performance on Test Set:"
```

	Actual	Prediction	LCB	UCB
2	16	23.344186	-82.63863	129.3270
3	11	8.662545	-107.47370	124.7988
5	28	-5.358733	-132.70435	121.9869
6	30	41.764009	-57.16067	140.6887
7	34	65.972495	-29.43544	161.3804
18	130	120.528880	22.97268	218.0851

```
[1] "VIF Analysis:"
```

	Year	Total.Operations	VFR.Total	IRF.Total
	6.595994	189.923163	91.212485	44.419993

Figure 3.5. Performance Analysis of Initial Class A SLR Model

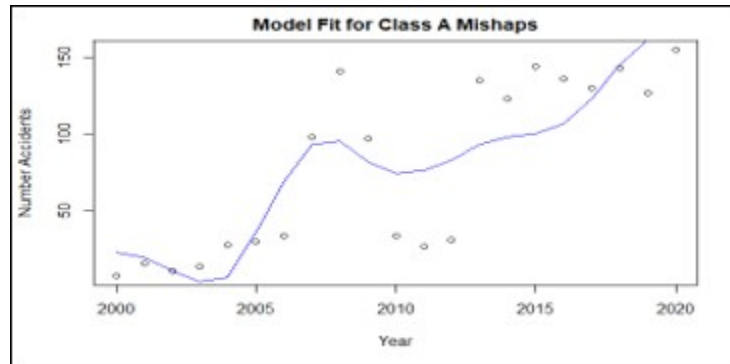


Figure 3.6. Fit of Initial Class A SLR Model to Actual Class A Occurrence

The initial model for predicting Class B mishaps was formulated as follows:

$$Y_B = \beta_0 + \beta_1 X_{Year} + \beta_2 X_{Total.Operations} + \beta_3 X_{VFR.Total} + \beta_4 X_{IFR.Total}$$

where:

- $Y_B \in [0, \infty)$ = the number of Class B mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = Year coefficient
- $X_{Year} = \langle 2000, \dots, 2020 \rangle$ = vector of Year predictors
- $\beta_2 \in (-\infty, \infty)$ = Total Operations coefficient
- $X_{Total.Operations}$ = vector of the number of Total Operations predictors for each year
- $\beta_3 \in (-\infty, \infty)$ = VFR Operations coefficient
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year
- $\beta_4 \in (-\infty, \infty)$ = IFR Operations coefficient
- $X_{IFR.Total}$ = vector of the number of IFR Operations predictors for each year.

The performance results of the initial Class B model show an incredibly large difference between the Training and Testing RMSE (Figure 3.7). Negative or skewed predictions for the number of Class B mishaps against the test data set's actual Y values reveals a model that is poorly fit to the data. VIF analysis reveals high multicollinearity between Total.Operations, VFR.Total, and IFR.Total predictors, though of a different interaction than was observed on the Class A model. As was seen with the initial Class A model, the initial Class B model does not conform well to the three modeling assumptions; transformations are required.

```
[1] "Training RMSE: 103.507262400115"
[1] "Testing RMSE: 81835.5348059343"
[1] "Performance on Test Set:"
```

	Actual	Prediction	LCB	UCB
2	33	136.17174	-205.6286	477.9720
3	50	-196.82455	-571.3702	177.7211
5	519	-76.31452	-487.0110	334.3820
6	743	514.40344	195.3660	833.4409
7	723	834.22878	526.5331	1141.9245
18	1102	1122.46560	807.8416	1437.0896

```
[1] "VIF Analysis:"
```

	Year	Total.Operations	VFR.Total	IFR.Total
	6.595994	189.923163	91.212485	44.419993

Figure 3.7. Performance Analysis of Initial Class B SLR Model

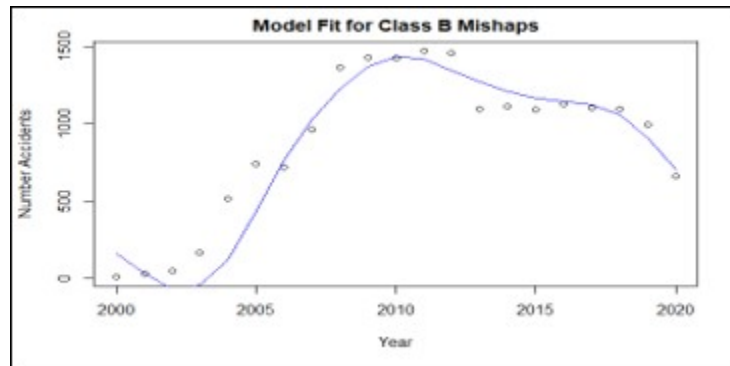


Figure 3.8. Fit of Initial Class B SLR Model to Actual Class B Occurrence

3.2.3 Multicollinearity Analysis

The results of the initial models necessitated an in-depth analysis of the multicollinearity interaction between the Total.Operations, VFR.Total, and IFR.Total predictors. Analysis revealed the following approximate relationships:

- A cubic relationship between Total.Operations and VFR.Total predictors
- An exponential (quadratic) relationship between Total.Operations and IFR.Total predictors
- An either exponential or cubic relationship between the VFR.Total and IFR.Total predictors

This analysis served as the basis for the transformed models, where VFR.Total and IFR.Total would need to be either poly(2) or poly(3) relationships and cross-factored. Additionally, it was concluded to remove the Total.Operations predictor altogether so as keep the number of predictors in the models as large as possible.

3.2.4 Predicting the Number of Mishaps — Intermediate Models (Transformations)

Many iterations of transformations were attempted with the objective of decreasing test RMSE and bringing the predicted Y values closer to their actual Y values. The final transformed Class A model was formulated as follows:

$$Y_A = \beta_0 + \beta_1 X_{Year} + \beta_2 (X_{VFR.Total} : X_{IFR.Total}^2)$$

where:

- $Y_A \in [0, \infty)$ = the number of Class A mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = Year coefficient
- $X_{Year} = \langle 2000, \dots, 2020 \rangle$ = vector of Year predictors
- $\beta_2 \in (-\infty, \infty)$ = coefficient for VFR Operations and IFR Operations crossfactor interaction
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year
- $X_{IFR.Total}$ = vector of the number of IFR Operations predictors for each year.

While the difference between training and test RMSE was somewhat decreased in this model, VIF was significantly reduced, indicating an elimination or near elimination of multicollinearity (Figure 3.9). This model did present the best transformation evolution in terms of fit and test RMSE value for Class A mishaps of the numerous evolutions attempted.

```
[1] "Training RMSE: 35.8134949610653"
[1] "Testing RMSE: 383.753083585135"
[1] "Performance on Test Set:"
  Actual Prediction      LCB      UCB
2      16    23.95506 -82.40821 130.3183
3       11    26.69368 -77.14310 130.5305
5       28    43.71341 -57.13558 144.5624
6       30    58.38760 -44.82944 161.6046
7       34    64.11077 -37.12951 165.3511
18      130   135.78951  35.67113 235.9079
[1] "VIF Analysis:"
              GVIF Df GVIF^(1/(2*Df))
Year              3.196352  1          1.787834
VFR.Total:poly(IFR.Total, 2) 3.196352  2          1.337099
```

Figure 3.9. Performance Analysis of Transformed Class A SLR Model

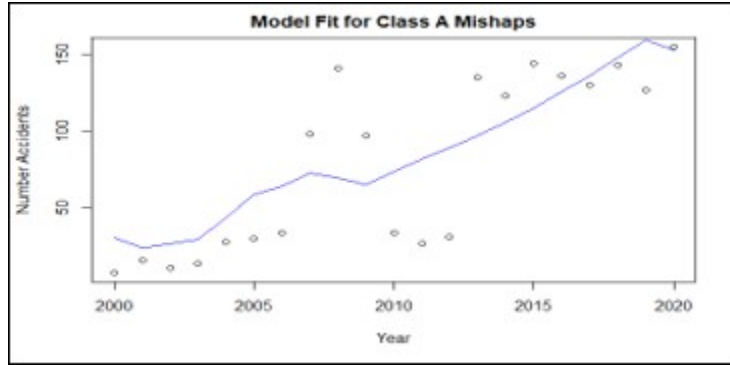


Figure 3.10. Fit of Transformed Class A SLR Model to Actual Class A Occurrence

Just as was the case with the intermediate Class A models, many iterations of transformations were attempted with the objective of decreasing test RMSE and bringing the predicted Y values closer to their actual Y values. The final transformed Class B model was formulated as follows:

$$Y_B = \beta_0 + \beta_1 X_{Year} + \beta_2 (X_{VFR.Total}^2 : X_{IFR.Total}^3)$$

where:

- $Y_B \in [0, \infty)$ = the number of Class B mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = Year coefficient
- $X_{Year} = \langle 2000, \dots, 2020 \rangle$ = vector of Year predictors
- $\beta_2 \in (-\infty, \infty)$ = coefficient for VFR Operations and IFR Operations crossfactor interaction
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year
- $X_{IFR.Total}$ = vector of the number of IFR Operations predictors for each year.

While the difference between training and test RMSE was somewhat decreased, VIF was significantly reduced, indicating an elimination or near elimination of multicollinearity (Figure 3.11). This model did present the best transformation evolution in terms of fit and test RMSE value for Class B mishaps of the numerous evolutions attempted. This model

would prove to be the best Class B SLR predictive model developed during POC1.

```
[1] "Training RMSE: 67.8925384083715"
[1] "Testing RMSE: 42973.5395140626"
[1] "Performance on Test Set:"
      Actual Prediction      LCB      UCB
2       33    294.1502 -266.6231  854.9235
3       50   -110.0536 -580.2172  360.1099
5      519   165.7316 -238.3412  569.8045
6      743   662.9098  259.0399 1066.7797
7      723   900.6505  561.7974 1239.5035
18     1102  1066.6168  807.4902 1325.7433
[1] "VIF Analysis:"
                                GVIF Df GVIF^(1/(2*Df))
Year                                4.14629  1          2.036244
poly(VFR.Total, 2):poly(IFR.Total, 3) 4.14629  6          1.125827
```

Figure 3.11. Performance Analysis of Transformed Class B SLR Model

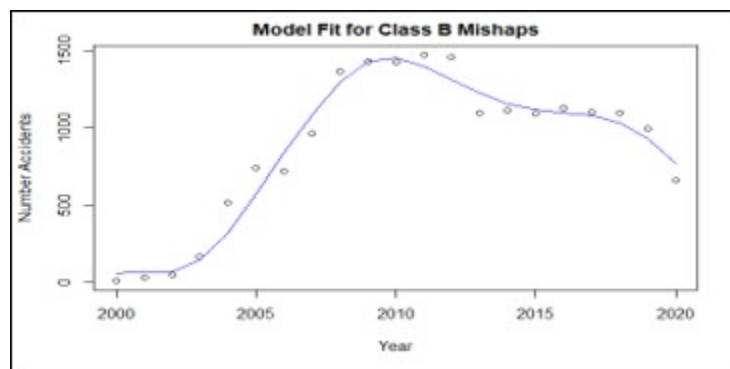


Figure 3.12. Fit of Transformed Class B SLR Model to Actual Class B Occurrence

Elimination of the multicollinearity issues indicated that the project could proceed into predictor selection to further refine the models.

3.2.5 Predictor Selection using Backward Step-wise AIC

Backward step-wise Akaike information criterion (AIC) was utilized to perform predictor selection on the Transformed models. Backward step-wise AIC was selected due to the size of the data set (< 50 predictors) and its preferred accuracy opposed to forward step-wise methodologies. Backward step-wise AIC determined the only useful predictor from the

Transformed Class A model was the Year predictor. Backward step-wise AIC determined that the Transformed Class B model was best as built.

3.2.6 Predictor Selection using Elastic Net Regularization

A cross-factored variation of the Class A Intermediate (Transformed) model predictors was fed into an elastic net regularization (ENR) model in an effort to further refine the model and to serve as a comparison to the AIC predictor selection. ENR is particularly advantageous in POC1 as the data includes both multicollinearity and sparsity, which ENR is specifically designed to handle.

ENR determined the best formulation for Class A prediction was as follows:

$$Y_A = \beta_0 + \beta_1 X_{VFR.Total}$$

where:

- $Y_A \in [0, \infty)$ = the number of Class A mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = Year coefficient
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year.

This model further reduced test RMSE compared to the Transformed Class A model (Figure 3.13), producing a better fit (Figure 3.14). This ENR model was determined to be the best Class A SLR prediction model produced by the project.

```
[1] "Training RMSE: 42.7974368148814"
[1] "Testing RMSE: 214.641349697466"
[1] "Performance on Test Set:"
```

	Actual	Prediction	LCB	UCB
2	16	23.31817	-94.636330	141.2727
3	11	18.01988	-102.143804	138.1836
5	28	28.21675	-87.803206	144.2367
6	30	45.61958	-64.446785	155.6859
7	34	58.91476	-47.677125	165.5066
18	130	112.10446	8.481643	215.7273

Figure 3.13. Performance Analysis of ENR-Informed Class A SLR Model

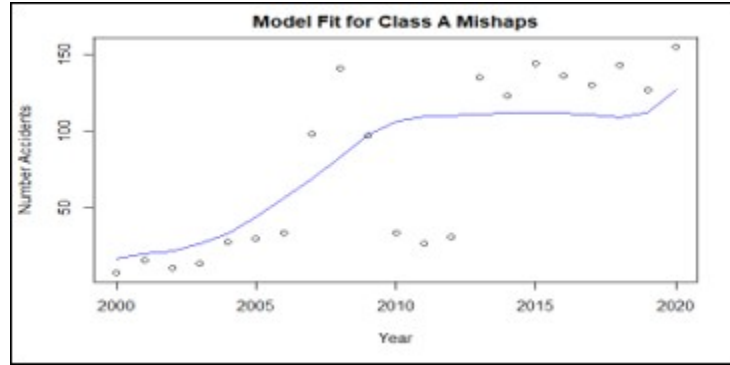


Figure 3.14. Fit of ENR-Informed Class A SLR Model to Actual Class A Occurrence

A cross-factored variation of the Class B Intermediate (Transformed) model predictors was fed into an ENR model to further refine the model and to serve as a comparison to the AIC predictor selection. However, this model performed very poorly, and thus the strict Class B Intermediate model was fed into ENR to attempt to further refine the model.

ENR determined the best formulation was as follows:

$$Y_B = \beta_0 + \beta_1(X_{VFR.Total}^2 : X_{IFR.Total}^3)$$

where:

- $Y_B \in [0, \infty)$ = the number of Class B mishaps
- $\beta_0 \in (-\infty, \infty)$ = intercept coefficient
- $\beta_1 \in (-\infty, \infty)$ = coefficient for VFR Operations and IFR Operations crossfactor interaction
- $X_{VFR.Total}$ = vector of the number of VFR Operations predictors for each year
- $X_{IFR.Total}$ = vector of the number of IFR Operations predictors for each year.

This model did not reduce test RMSE compared to the Transformed Class B model (Figure 3.15) and was a comparatively worse fit (Figure 3.16).

```
[1] "Training RMSE: 140.60563421541"
[1] "Testing RMSE: 34857.4675078298"
[1] "Performance on Test Set:"
```

	Actual	Prediction	LCB	UCB
2	33	304.6104	-754.7692	1363.990
3	50	-108.3101	-996.5511	779.931
5	519	260.9577	-497.2593	1019.175
6	743	662.3615	-100.6376	1425.361
7	723	879.4930	239.6280	1519.358
18	1102	1214.8365	745.1413	1684.532

Figure 3.15. Performance Analysis of ENR-Informed Class B SLR Model

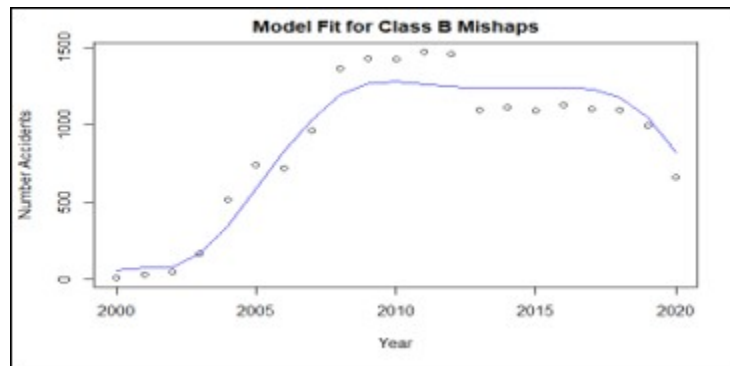


Figure 3.16. Fit of ENR-Informed Class B SLR Model to Actual Class B Occurrence

3.2.7 Comparison of Transformed Model and ENR Model Performance

A final performance comparison of the Intermediate (Transformed) models and the ENR models was undertaken. Models that produced smaller Confidence and Prediction intervals with more realistic means would be preferred. The 90% Confidence and Prediction Intervals were taken. As input data, the mean number of VFR and IFR flights were taken from the aggregate data set to serve as the projected number of VFR and IFR flights for year 2022.

For Class A mishap prediction, the ENR model (Figure 3.18) was confirmed as the better model over the Transformed model (Figure 3.17). The ENR model produced a better (more realistic) median fit and narrower intervals.

```
[1] "Intermediate (Transformed) Model 90% Confidence Interval:"
      fit      lwr      upr
187.5038 116.5086 258.4991
[1] "Intermediate (Transformed) Model 90% Prediction Interval:"
      fit      lwr      upr
187.50384  84.15376 290.85392
```

Figure 3.17. 90% Confidence and Prediction Intervals of Transformed Class A SLR Model

```
[1] "ENR Model 90% Confidence Interval:"
      fit      lwr      upr
80.91398  58.06771 103.76026
[1] "ENR Model 90% Prediction Interval:"
      fit      lwr      upr
80.913985 -3.643891 165.471860
```

Figure 3.18. 90% Confidence and Prediction Intervals of ENR-Informed Class A SLR Model

For Class B mishap prediction, the Transformed model (Figure 3.19) was confirmed as the better model over the ENR model (Figure 3.20). The Transformed model produced a better (more realistic) median fit and narrower intervals.

```
[1] "Intermediate (Transformed) Model 90% Confidence Interval:"
      fit      lwr      upr
802.3385  540.4400 1064.2369
[1] "Intermediate (Transformed) Model 90% Prediction Interval:"
      fit      lwr      upr
802.3385  479.7790 1124.8979
```

Figure 3.19. 90% Confidence and Prediction Intervals of Transformed Class B SLR Model

```
[1] "ENR Model 90% Confidence Interval:"
      fit      lwr      upr
1312.0762  993.4414 1630.7109
[1] "ENR Model 90% Prediction Interval:"
      fit      lwr      upr
1312.076  832.797 1791.355
```

Figure 3.20. 90% Confidence and Prediction Intervals of ENR-Informed Class B SLR Model

3.2.8 POC1 Conclusions

The best-performing SLR models for each class of mishap were as follows:

- Class A ENR Model
- Class B Intermediate (Transformed) Model

While the selected models perform decently well considering the sparsity of the data, they contain very wide Confidence and Prediction Intervals. Additional predictors are required to produce more accurate SLR models. However, the POC accomplished its primary objective of proving the feasibility of whether the number of Class A and B mishaps can be predicted.

The intent of this POC was to serve as an initial proof of concept into whether data modeling and predictive analysis can serve as viable tools to assist in the construction of targeted safety programs based on statistical analysis. To this end, POC1 pursued two objectives: determine the viability of predicting the number of Class A and B mishaps for a year given historical data; and if viability was proven, to determine which predictors were best suited to making this prediction.

Enumerated Answers to POC1 Questions:

1. It is possible to predict the number of Aviation Class A and B mishaps for a given year, but additional predictors are required to formulate more accurate SLR models than those constructed by this POC.
2. For the models constructed, the total number of VFR flights in a given year was the best predictor for Class A mishap prediction. The numbers of VFR and IFR flights operated in a given year were the best predictors for Class B mishap prediction. It must be noted, however, that tracking of these kinds of total flight subcategories is not captured in USAF data.

POC1 has concluded that accomplishment of the stated objectives with the desired end state is possible with additional input predictor data and SLR model refinement. However, additional predictor data is difficult to access from disparate USAF major commands. For example, if an additional predictor of Total Active Pilots were added to the model, such data would have to be pulled regularly by AFSEC from AFPC for any SLR predictive models implemented utilizing that data. Therefore, different predictive modeling methodologies

may be better suited to mishap data than SLR models.

3.3 Proof of Concept 2: Time Series Analysis

The conclusions from POC1 showed us that SLR models were perhaps not the best predictive modeling methodology to implement given individual mishap data sets. This caused the research team to examine whether TSA models would be better suited, considering the format of individual mishap data, which TSA modeling methodologies are built to utilize.

POC2, conducted across August-September 2021, sought to answer the following questions:

1. Are TSA predictive models better suited to predicting mishap incidence than SLR models?
2. Is there an identifiable trend for the occurrence of Aviation Class A and B mishaps?
3. What TSA model best predicts the number of Aviation Class A and B mishaps for a projected year?

As was the case in POC1, POC2 explored these research questions simultaneously through construction of two different models for Class A and B mishaps. POC2 utilized the same NTSB mishap data as used in POC1, which is similar in format and structure to the AFSAS data combined with FAA total operations data. POC2 analysis and modeling was performed using R.

3.3.1 Cleaning and Transforming Individual Mishap Data to TSA Format

Cleaning data is one of the most important processes a data analyst can perform, as the quality of the cleaning performed will determine the quality of the models built from the data. Cleaning is especially crucial for TSA models, which are dependent on a continuous time sequence of data. Therefore, any data being used in TSA must be sorted in chronological order and have an unbroken time sequence. Even if no events occurred for a given time period of the data, zeros must be entered for these periods. These principles drove the cleaning techniques of the data for POC2.

A benefit of TSA is that it can utilize raw incident data, where each entry or line in the

data set represents a single incident. This is not the case in SLR, where the data had to be transformed using Python Pandas prior to modeling. After the raw NTSB mishap data was loaded into R, the individual mishap Event Dates were converted to the R `date` data type. All columns containing categorical data were converted to the R `factor` data type. Any NA entries in numerical data columns were converted to zeros. The raw mishap data was then split between raw Class A and Class B data sets for counting.

To create the continuous, unbroken reference time series required by TSA models, a count of all incidents was performed for each month in the data set. The volume of the data ensured that no months were missing in the data; there were no months where a mishap of either class did not occur. However, the cleaning process was still coded such that each month was generated and accounted for between the start and end of the data set period while calculating the monthly counts of mishaps. The same process was then employed for each mishap class independently.

Once the total monthly counts, total Class A monthly counts, and total Class B monthly counts were calculated, the Class A and B monthly counts were left-hand merged onto the total monthly counts data frame by date. Merging the classes onto the monthly counts data frame ensured that the continuity of the data set period would be maintained if there were any months where either Class A or Class B mishaps did not occur. Any NA entries in the counts left after the merge for either Class A or B mishaps were filled with zeros. This resulting data frame would serve as the primary data frame of reference for the fitting of TSA models.

3.3.2 Fitting Seasonal Decomposition Models

The Seasonal Decomposition of Class A mishaps was the first TSA performed. The decomposition was derived through use of R's `ts()` function with a period frequency of 12, meaning the expected seasonality was to be one year long in duration. A periodic Seasonal Decomposition model was then fitted to the Class A decomposition using R's `stl()` function. The Seasonality, Trend, and Remainder component values were also extracted for examination.

Examination of the time-series decomposition for monthly Class A mishaps revealed a strong seasonal trend (Figure 3.21). Class A incidents peak in July each year and ebb in

February each year. The Class A mishap trend has been increasing since July 2013 (Figure 3.22). There was a large amount of noise (randomness) present in the data as can be observed in the “Remainder” of Figure 3.21. This was expected given the inherently random nature of mishap occurrence. This magnitude of noise is also a good thing: while a small amount of randomness would make predicting on this data easier, it would also indicate strong systemic issues correlated with the causes of mishap occurrence.

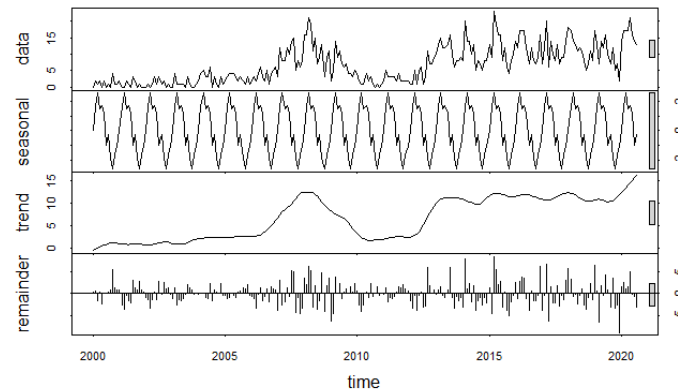


Figure 3.21. NTSB Class A Mishap Seasonal Decomposition

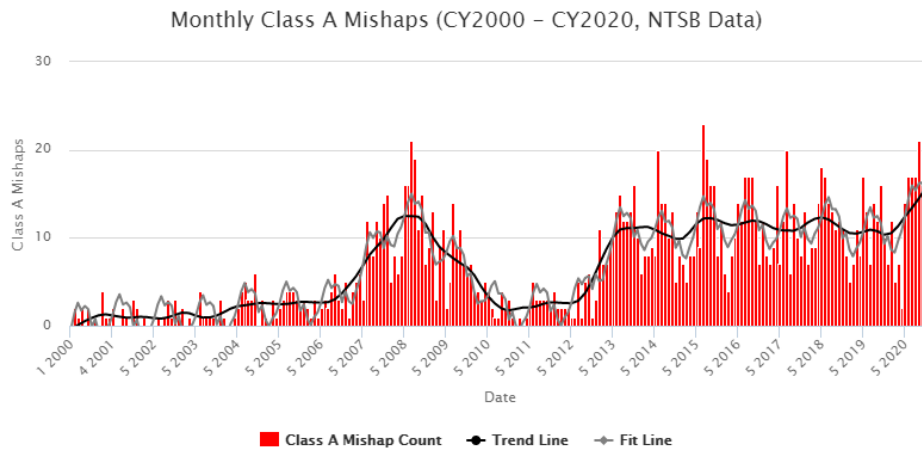


Figure 3.22. NTSB Class A Mishap Seasonality and Trend Plotted over Monthly Counts

The Seasonal Decomposition of Class B mishaps was extracted and modeled utilizing

the same methodology described for the Class A mishaps. Examination of the time-series decomposition for monthly Class B mishaps reveals a strong seasonal trend (Figure 3.23). Class B incidents peak in July each year and ebb in January each year. The Class B mishap trend has been decreasing since July 2018 (Figure 3.24). A large amount of randomness is also present in the Class B mishap decomposition.

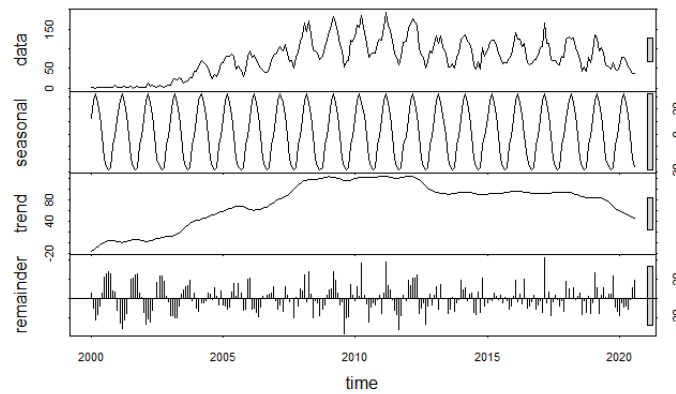


Figure 3.23. NTSB Class B Mishap Seasonal Decomposition

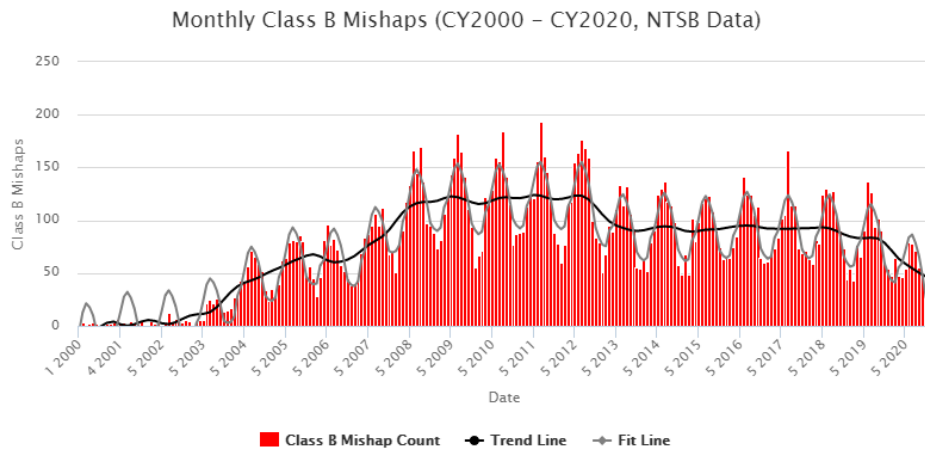


Figure 3.24. NTSB Class B Mishap Seasonality and Trend Plotted over Monthly Counts

3.3.3 Fitting TSA Predictive Models

To determine which TSA models best predict the Class A mishap incidence, we fit Naive, Holt-Winters, Naive Seasonal Decomposition, and ARIMA predictive models to the NTSB Class A monthly counts. We then performed one-step ahead rolling horizon predictions for the next period in the models, in this case the next 12 months. With these predictions, we were able to measure the performance of each model by calculating their MAPE and MASE values. We calculated one-, three-, six-, and 12-month prediction MAPE and MASE values for each model, with a focus on 12-month predictions. This process was performed independently for Class A and Class B mishaps to determine which models would be better used for predicting each phenomenon. Long-term predictive performance was preferred for this model, as the intent for these models is to predict mishap occurrences by month for the next FY for safety planners. Therefore, the 12-step MAPE and MASE values were primarily used for evaluation.

Examination of the MAPE and MASE values for all Class A mishap TSA predictive models revealed that Seasonal Decomposition and ARIMA models perform the same when examining their long-term (12-month) MAPE performance (Figure 3.25). Seasonal Decomposition outperformed the ARIMA model in long-term MASE performance by a small margin. Therefore, we examined both Seasonal Decomposition and ARIMA for in-depth evaluation of predictive performance on Class A mishaps.

Model	1 Step MAPE	3 Step MAPE	6 Step MAPE	12 Step MAPE	1 Step MASE	3 Step MASE	6 Step MASE	12 Step MASE
Naive	0.77	0.77	0.78	0.77	1	1	1	1
Seasonal Decomposition	0.6	0.58	0.74	0.71	0.79	0.83	1	0.96
Holt-Winters	0.52	0.62	0.76	0.72	0.73	0.95	1.1	0.99
ARIMA	0.52	0.6	0.77	0.71	0.74	0.93	1.1	0.98

Figure 3.25. NTSB Class A Mishap TSA Model Performance (MAPE & MASE) under One-Step Forward Horizon Predictions

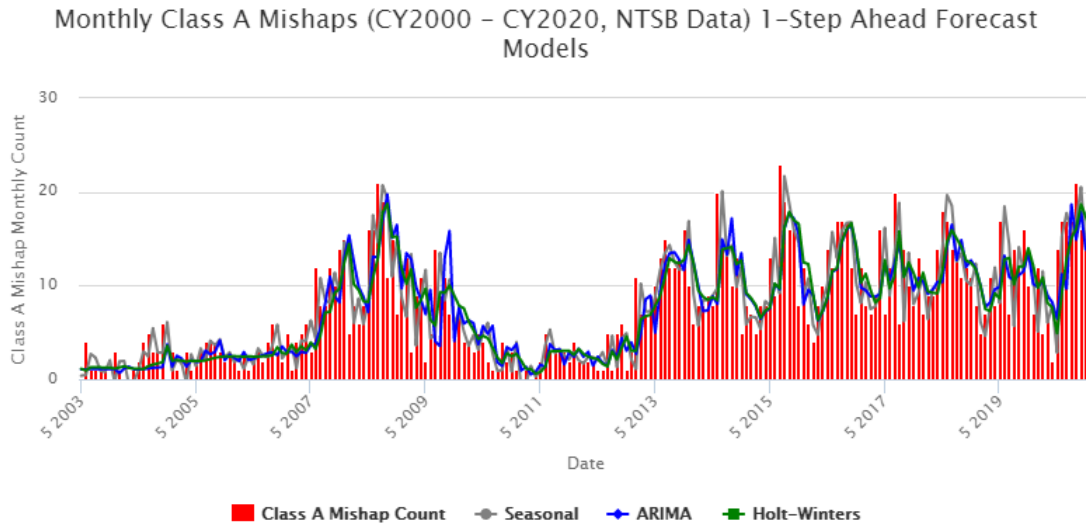


Figure 3.26. NTSB Class A TSA Fitted Models Plotted over Monthly Counts

Examination of the MAPE and MASE values for all Class B mishap TSA predictive models revealed that Seasonal Decomposition and ARIMA models performed the same when examining their long-term (12-month) MAPE performance (Figure 3.27). In this case, opposite of Class A mishaps, ARIMA outperformed the Seasonal Decomposition model in long-term MASE performance by a small margin. Therefore, we examined both Seasonal Decomposition and ARIMA for in-depth evaluation of predictive performance on Class B mishaps.

Model	1 Step MAPE	3 Step MAPE	6 Step MAPE	12 Step MAPE	1 Step MASE	3 Step MASE	6 Step MASE	12 Step MASE
Naive	0.24	0.23	0.23	0.23	1	1	1	1
Seasonal Decomposition	0.17	0.24	0.3	0.25	0.77	1.1	1.3	1.1
Holt-Winters	0.21	0.44	0.63	0.44	0.97	2	2.8	2
ARIMA	0.16	0.22	0.27	0.25	0.72	0.91	1.1	1

Figure 3.27. NTSB Class B Mishap TSA Model Performance (MAPE & MASE) under One-Step Forward Horizon Predictions

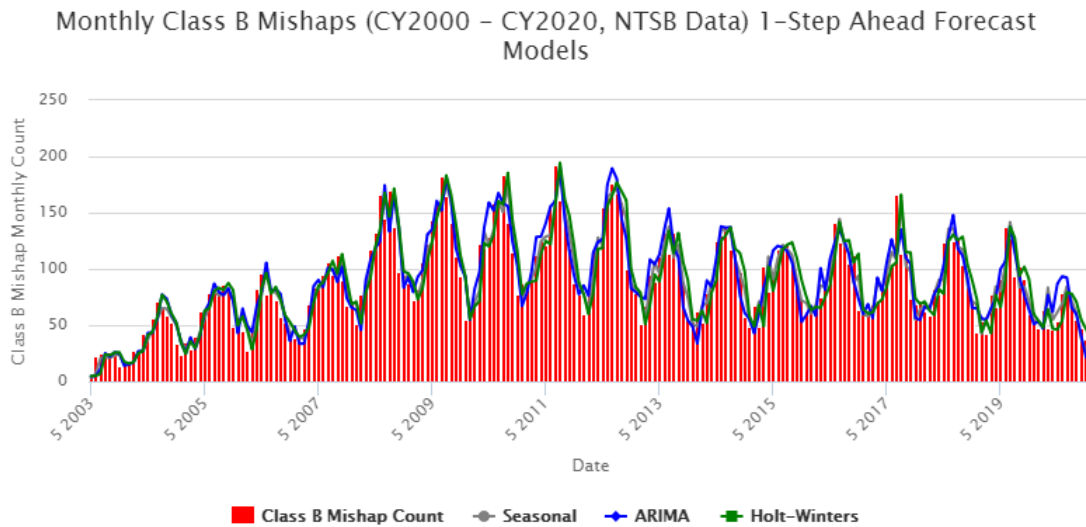


Figure 3.28. NTSB Class B TSA Fitted Models Plotted over Monthly Counts

3.3.4 Seasonal Decomposition Model Forecasting

In-depth examination of model prediction intervals is not only an indicator of model predictive performance, but is also a gauge into the usefulness of the model for making predictions in real-world situations. Wider prediction intervals indicate less-accurate predictions, which could invalidate the model's usefulness in real-world application if the intervals are too wide. Narrower prediction intervals indicate more-accurate predictions, and can indicate usefulness to safety professionals in predicting mishap incidence for safety planning.

We began the in-depth examination of predictive performance by fitting and then forecasting with a Naive Seasonal Decomposition model for both Class A and B mishaps. We predicted the next period of data, or 12 months, and examined the width of the 80% and 90% Prediction Intervals to gauge performance of the model. Predictions were generated using R's `predict()` function utilizing the Naive Seasonal Decomposition model generated in Section 3.3.2.

Results of the Class A Naive Seasonal Decomposition prediction intervals are shown in Figures 3.29 and 3.30.

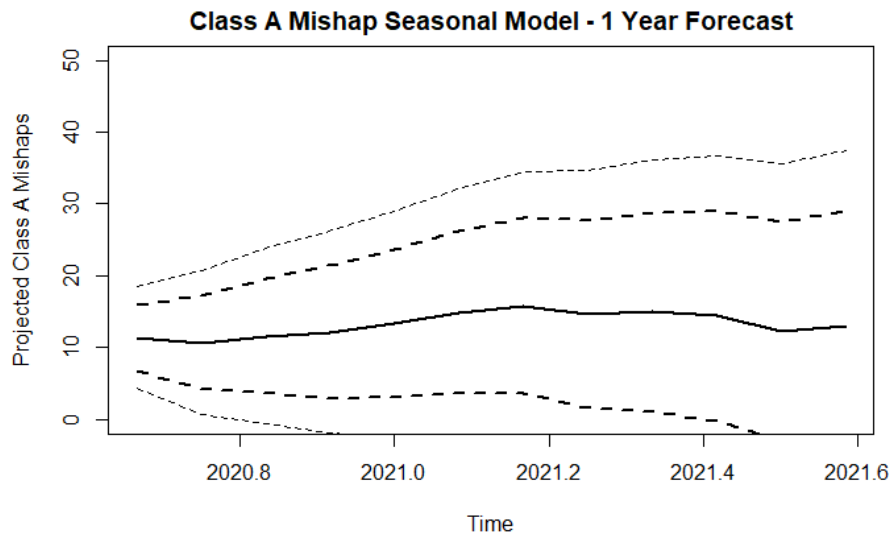


Figure 3.29. NTSB Class A Seasonal Decomposition Model Prediction with 80% and 90% Prediction Intervals

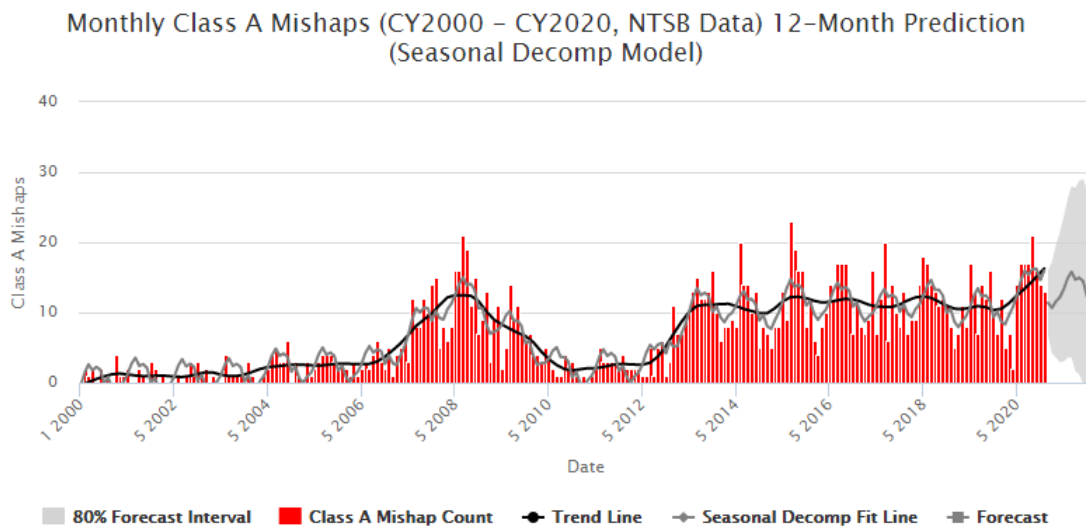


Figure 3.30. NTSB Class A Seasonal Decomposition Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Results of the Class B Naive Seasonal Decomposition prediction intervals are shown in Figures 3.31 and 3.32.

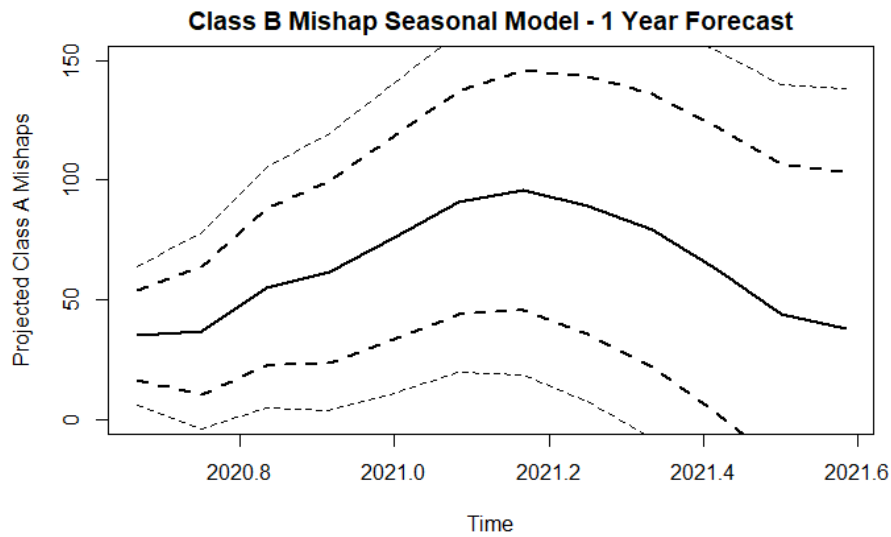


Figure 3.31. NTSB Class B Seasonal Decomposition Model Prediction with 80% and 90% Prediction Intervals

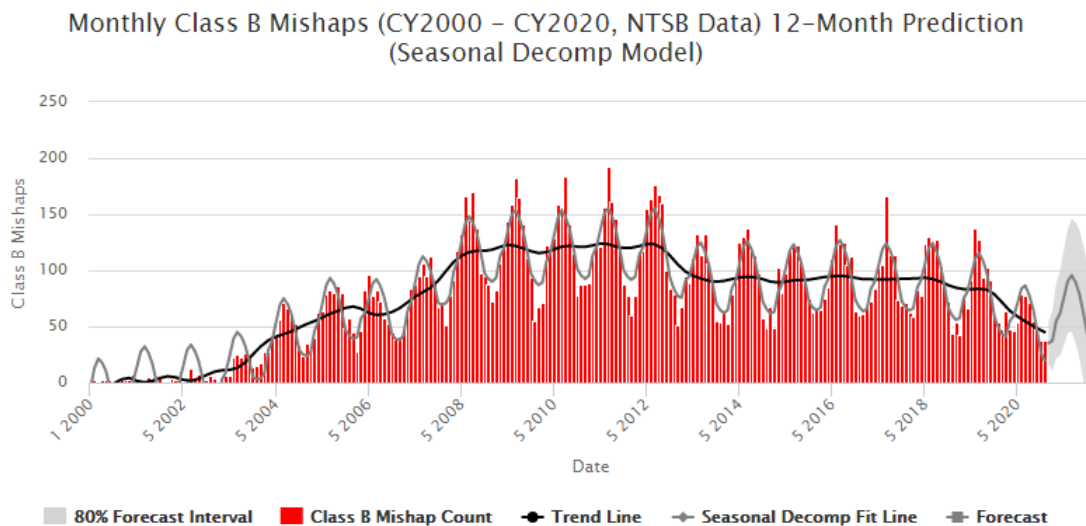


Figure 3.32. NTSB Class B Seasonal Decomposition Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Examination of the Class A and B 1-year Naive Seasonal Decomposition forecasts reveal extremely wide 80% and 90% Confidence Intervals, with growing uncertainty through the

projected year. Despite the wide prediction intervals for Class A mishaps, the Seasonal Decomposition model could be extremely useful for aviation safety professionals. This model will afford safety planners insight to the projected peak and ebb months for the next year. Additionally, the number of mishaps projected can provide measure of performance (MOP)s for safety program effectiveness: if the year progresses below the forecast, the program may have been successful at reducing mishaps and the best practices for the plan should be identified; if the year progresses above the forecast, the program may not have been successful at reducing mishaps and requires improvement.

3.3.5 ARIMA Predictive Model Forecasting

We next performed the in-depth examination of predictive performance by fitting and then forecasting with a seasonal ARIMA model for both Class A and B mishaps utilizing R's `auto.arima()` function. We hoped that by including seasonality into the ARIMA model we would produce a better-informed model than the Naive Seasonal Decomposition model. We predicted the next period of data, or 12 months, and examined the width of the 80% and 90% Prediction Intervals to gauge performance of the model. Predictions were generated using R's `forecast()` function utilizing a Seasonal ARIMA model fitted against the time-series decomposition generated in Section 3.3.2.

Results of the Class A Seasonal ARIMA prediction intervals are shown in Figures 3.33 and 3.34.

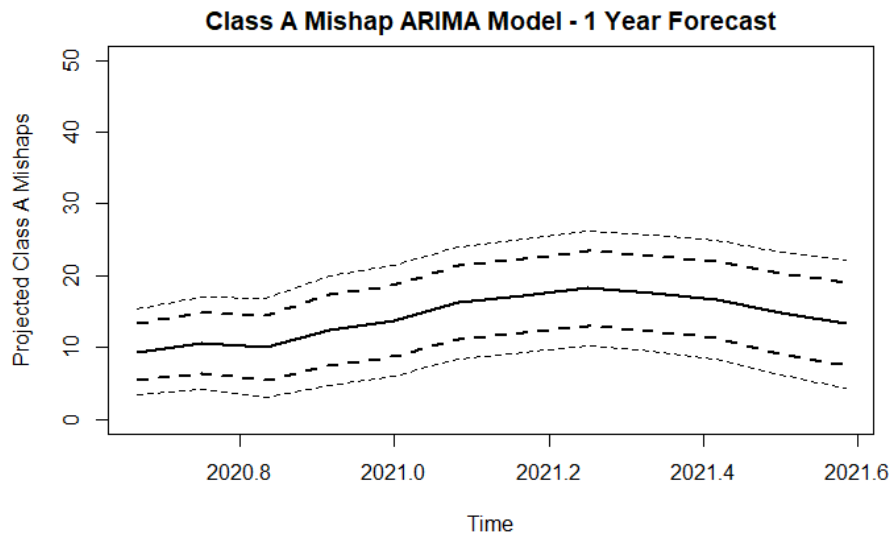


Figure 3.33. NTSB Class A Seasonal ARIMA Model Prediction with 80% and 90% Prediction Intervals

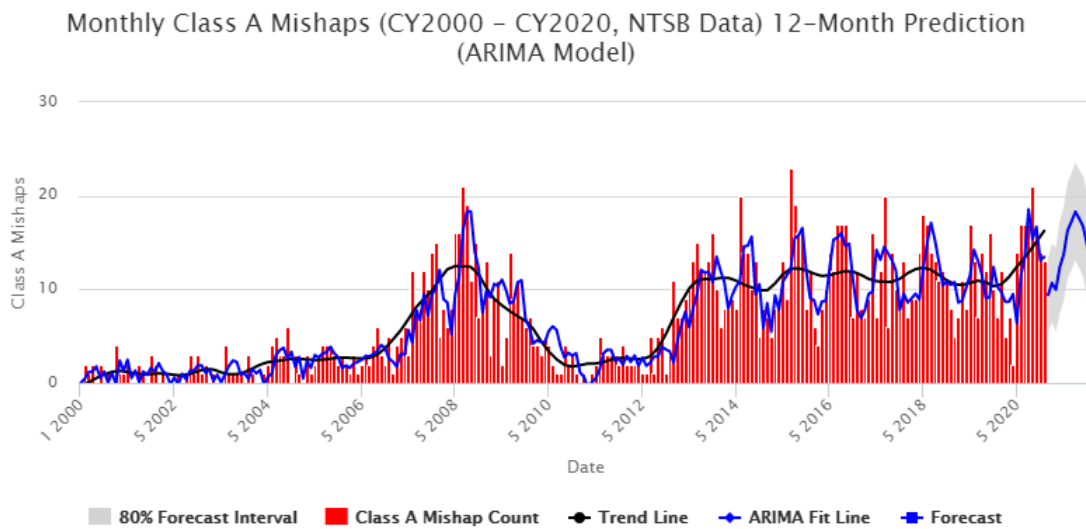


Figure 3.34. NTSB Class A Seasonal ARIMA Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Results of the Class B Seasonal ARIMA prediction intervals are shown in Figures 3.35 and 3.36.

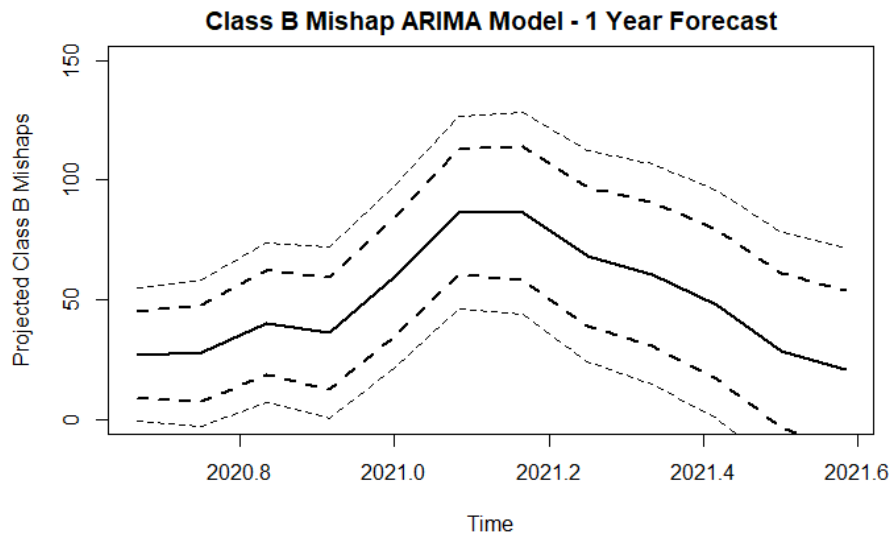


Figure 3.35. NTSB Class B Seasonal ARIMA Model Prediction with 80% and 90% Prediction Intervals

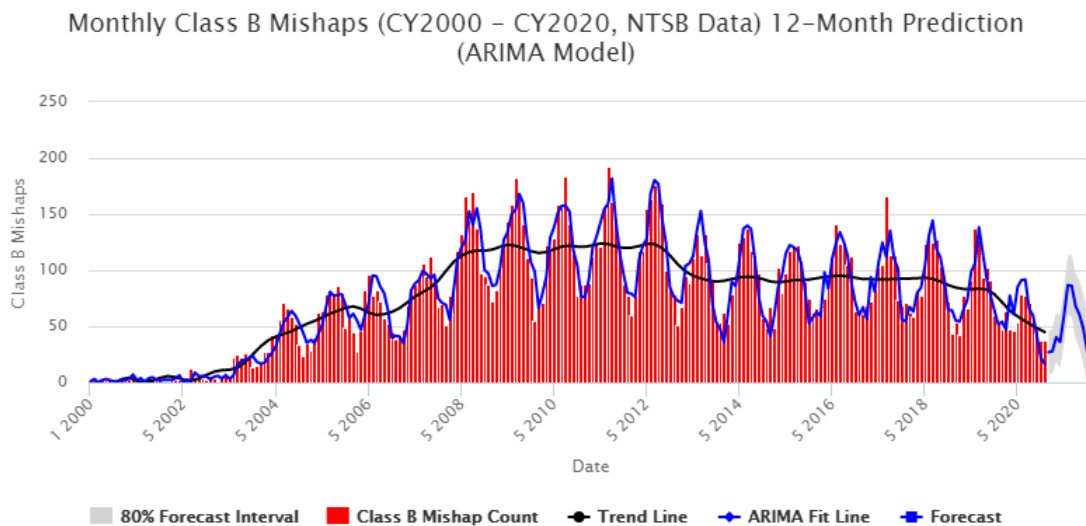


Figure 3.36. NTSB Class B Seasonal ARIMA Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Examination of the Class A and B 1-year Seasonal ARIMA forecasts reveal wide 80% and 90% confidence intervals, with growing uncertainty through the projected year. The

confidence intervals produced by the Seasonal ARIMA model are narrower than those generated by the Seasonal Decomposition, however. This is a result of the moving average that is a component of the ARIMA model. The narrower prediction intervals with the Seasonal ARIMA model makes this model extremely useful for aviation safety professionals. This model will afford safety planners insight to the projected average normal number of mishaps for the next year. As was the case with the Seasonal Decomposition model, the number of mishaps projected by the Seasonal ARIMA model can provide MOPs for safety program effectiveness within the context of a “Band of Normal”: if the number of mishaps progresses below the prediction interval, the number of mishaps observed are below the normal range of mishaps observed. This should alert safety professionals to further analyze why mishaps are dipping below the normal range of occurrence. If the number of mishaps progresses above the prediction interval, the number of mishaps observed are above the normal range of mishaps. This should alert safety professionals to further analyze why mishaps are trending over the normal range of occurrence.

3.3.6 POC2 Conclusions

Enumerated Answers to POC2 Questions:

1. TSA models are a far superior predictive modeling methodology for Aviation mishap incidence given the format and quality of mishap data.
2. Our research reveals that there is an identifiable trend for the incidence of Class A and B mishaps in the NTSB data. There is also strong seasonality for both Class A and Class B mishaps. Class A mishaps peak in July and ebb in February each year. This seasonality establishes an incidence baseline of an average of 14 Class A mishaps at peak and 8 Class A mishaps at the low point. Class B mishaps peak in July and ebb in January each year. This establishes a Class B incidence baseline of an average of 120 at peak and 60 at low point.
3. For Class A mishaps, the ARIMA model fits slightly better than the Seasonal Decomposition model. For Class B mishaps, the opposite is true, with the Seasonal Decomposition model fitting slightly better than the ARIMA model. Given this better fit, the determination between which model is better depends on the preferred prediction task. The Seasonal Decomposition model better incorporates the seasonal nature and randomness inherent in Aviation mishap incidence, and therefore includes wider

confidence intervals. The ARIMA model, however, better closely follows the overall trend of the data set, and includes tighter confidence intervals.

POC2 demonstrated the ability to apply TSA models to aviation mishap incident data with decent predictive power. POC2 laid the foundation for analysis of the AFSAS data, to which the POC2 code would be applied once the data was received by the research team.

3.4 Applying Proof of Concept 2 to AFSAS Data

The research team received the AFSAS data in September 2021. The code from POC2 was applied to the data set, with the research team modifying the code as required to fit the AFSAS data. The following analysis will be generalized, as it is meant to demonstrate that the methods developed in POC2 were simply applied to the AFSAS data; in-depth discussion of the findings will be discussed in Chapter 4.

3.4.1 Cleaning and Transforming AFSAS Individual Mishap Data to TSA Format

Minimal changes were required from the POC2 methodology of importing and cleaning the data, as the only changes required were to account for the different structure of the AFSAS data in comparison to the NTSB data. Therefore, only the number and identity of data columns converted to date and factor data types changed. There were no required changes to the counting methodology as utilized in POC2.

3.4.2 Fitting Seasonal Decomposition Models to AFSAS Data

No changes were required to the POC2 time-series decomposition methodology when applied to the AFSAS data.

Examination of the time-series decomposition for monthly Class A mishaps (Figure 3.37) revealed a strong seasonal trend. There was an even larger amount of noise (randomness) present in the AFSAS data than was observed in the NTSB data. This was not surprising, given the low number of Class A mishaps that occur in military aviation. This magnitude of noise is also a good thing; while a small amount of randomness would make predicting

on this data easier, it would also indicate strong systemic issues correlated with the causes of mishap occurrence.

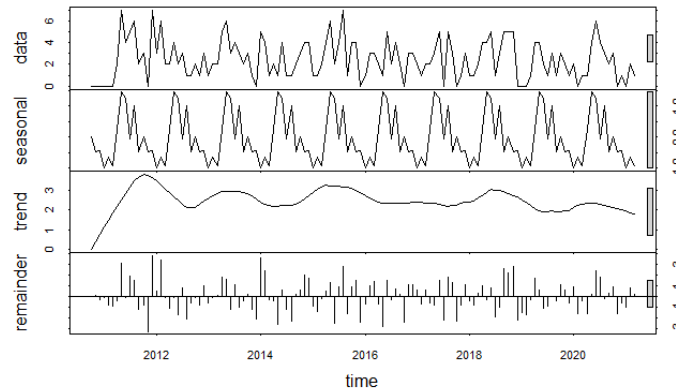


Figure 3.37. AFSAS Class A Mishap Seasonal Decomposition

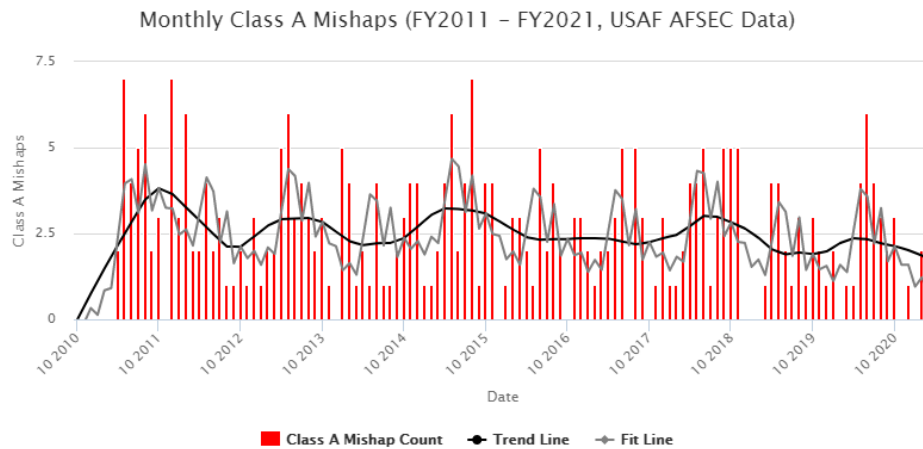


Figure 3.38. AFSAS Class A Mishap Seasonality and Trend Plotted over Monthly Counts

Examination of the time-series decomposition for monthly Class B mishaps (Figure 3.39) revealed a strong seasonal trend. A similar increase in the amount of randomness that was observed with the Class A data is also present in the Class B mishap decomposition.

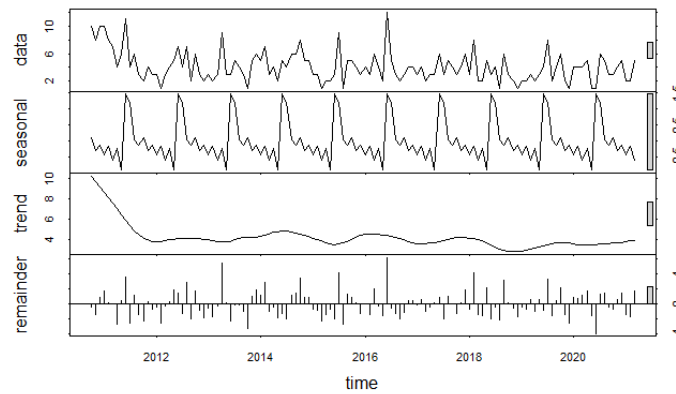


Figure 3.39. AFSAS Class B Mishap Seasonal Decomposition

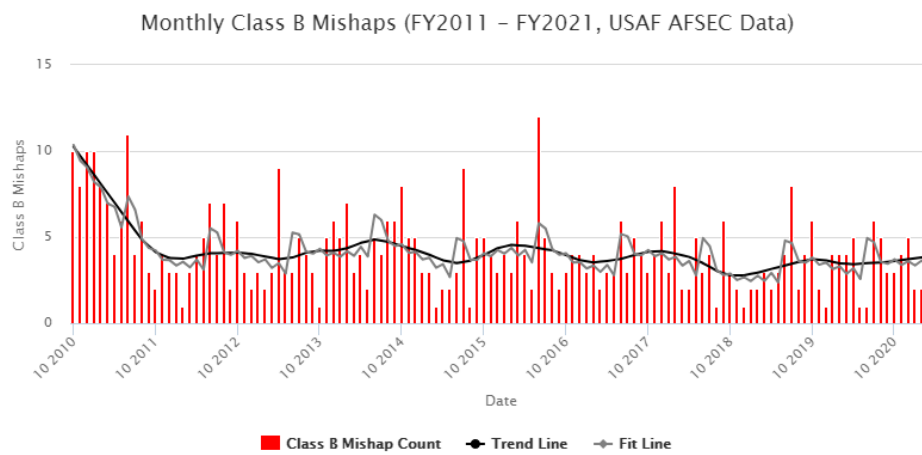


Figure 3.40. AFSAS Class B Mishap Seasonality and Trend Plotted over Monthly Counts

3.4.3 Fitting TSA Predictive Models to AFSAS Data

No changes were required to the POC2 methodology of fitting TSA models when applied to the AFSAS data. Given the strong performance and potential uses of both Seasonal Decomposition and ARIMA models in POC2, the performance of these models when applied to the AFSAS data was of particular interest.

Examination of the MAPE and MASE values for all Class A mishap TSA predictive models

revealed that the ARIMA model performed the best with respect to long-term (12-month) MAPE and MASE performance (Figure 3.41). It was also noted that the MAPE and MASE values for all models was much higher when applied to the AFSAS data than what was observed with the NTSB data. This is due to the smaller number of mishaps present in the AFSAS data, which encompasses only the fleet of the USAF, when compared to the U.S.-wide domain of the NTSB. It is always inherently more difficult to predict accurately off smaller data sets.

Model	1 Step MAPE	3 Step MAPE	6 Step MAPE	12 Step MAPE	1 Step MASE	3 Step MASE	6 Step MASE	12 Step MASE
Naive	1.1	1.1	1	1	1	1	1	1
Seasonal Decomposition	1.2	1.2	1	1.1	1	1	1	1.1
Holt-Winters	1	1	0.98	0.98	0.91	0.88	0.91	0.91
ARIMA	0.92	0.87	0.84	0.88	0.82	0.79	0.81	0.82

Figure 3.41. AFSAS Class A Mishap TSA Model Performance (MAPE & MASE) under One-Step Forward Horizon Predictions

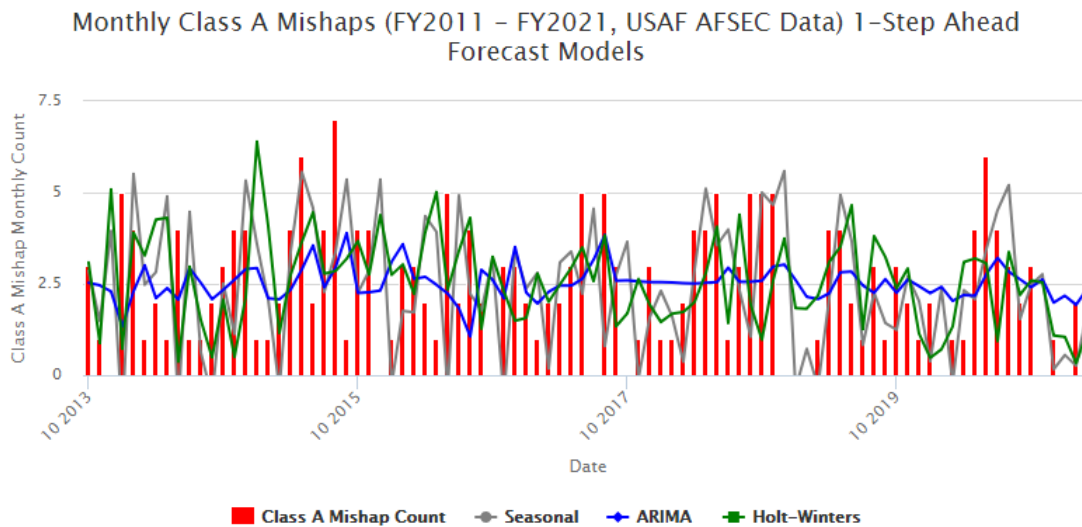


Figure 3.42. AFSAS Class A TSA Fitted Models Plotted over Monthly Counts

Examination of the MAPE and MASE values for all Class B mishap TSA predictive models revealed that the ARIMA model performed the best when examining long-term (12-month)

MAPE and MASE performance (Figure 3.43).

Model	1 Step MAPE	3 Step MAPE	6 Step MAPE	12 Step MAPE	1 Step MASE	3 Step MASE	6 Step MASE	12 Step MASE
Naive	0.74	0.73	0.73	0.73	1	1	1	1
Seasonal Decomposition	0.76	0.84	0.78	0.79	1	1	1	1
Holt-Winters	0.67	0.75	0.73	0.73	0.92	0.99	0.97	0.99
ARIMA	0.61	0.63	0.65	0.64	0.76	0.78	0.8	0.79

Figure 3.43. AFSAS Class B Mishap TSA Model Performance (MAPE & MASE) under One-Step Forward Horizon Predictions

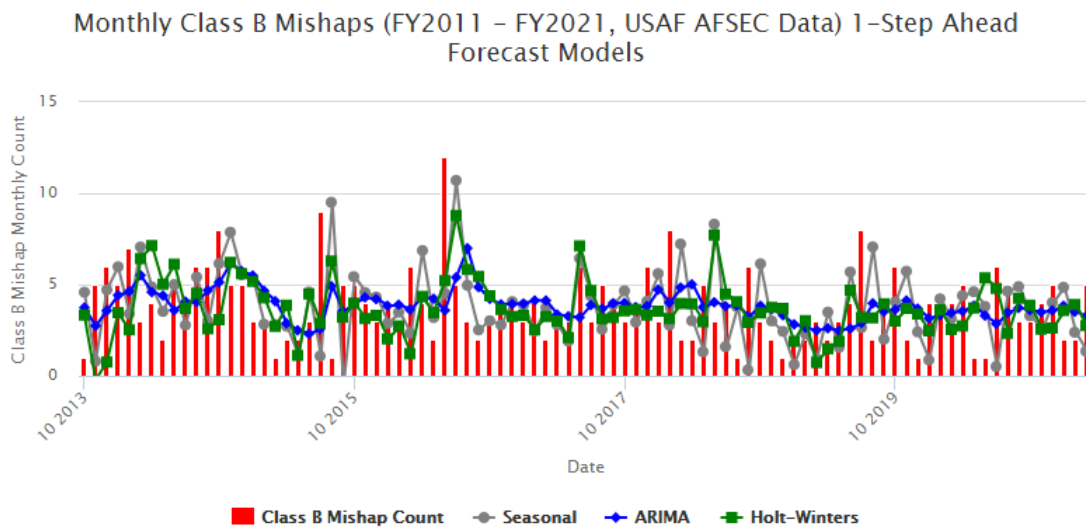


Figure 3.44. AFSAS Class B TSA Fitted Models Plotted over Monthly Counts

While the ARIMA models outperformed all other TSA models for both Class A and B USAF mishaps, it must be noted that the long-term performance was not much better over the performance of the Seasonal Decomposition models. The research team felt the weaker predictive performance did not negate the proposed use case for the Seasonal Decomposition model.

3.4.4 Seasonal Decomposition Model Forecasting for AFSAS Data

The methodology for the in-depth examination of the Seasonal Decomposition model forecast changed only slightly from POC2. Only the 80% Confidence Intervals around the prediction were examined with respect to the overall model and historical mishap counts.

Results of the Class A Naive Seasonal Decomposition prediction intervals are shown in Figure 3.45.

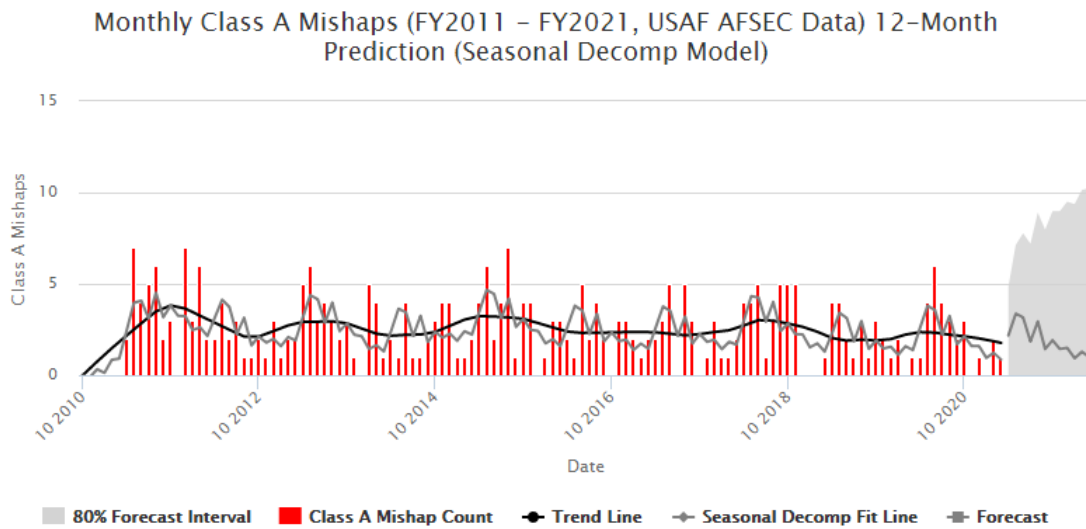


Figure 3.45. AFSAS Class A Seasonal Decomposition Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Results of the Class B Naive Seasonal Decomposition prediction intervals are shown in Figure 3.46.

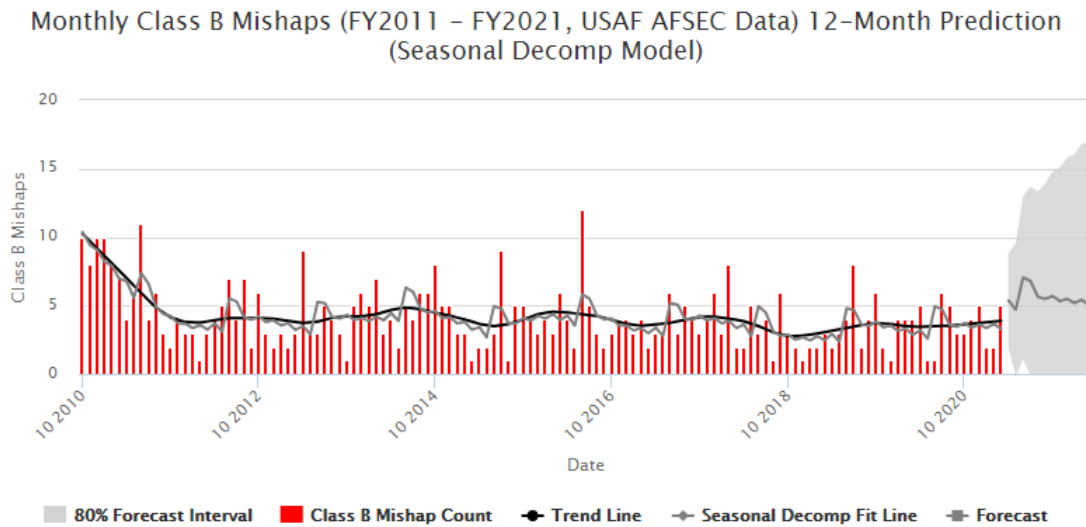


Figure 3.46. AFSAS Class B Seasonal Decomposition Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Examination of the Class A and B 1-year Naive Seasonal Decomposition forecasts reveal extremely wide 80% Confidence Intervals with respect to the overall historical monthly counts, with growing uncertainty through the projected year. Despite the wide prediction interval for Class A mishaps, the Seasonal Decomposition model could be extremely useful for aviation safety professionals by providing the MOPs discussed in Section 3.3.4. See Section 3.5.2 for further details on the use case for the Seasonal Decomposition model.

3.4.5 ARIMA Predictive Model Forecasting for AFSAS Data

The methodology for the in-depth examination of the ARIMA model forecast changed only slightly from POC2. Only the 80% Confidence Intervals around the prediction were examined with respect to the overall model and historical mishap counts.

Results of the Class A Seasonal ARIMA prediction intervals are shown in Figure 3.47.

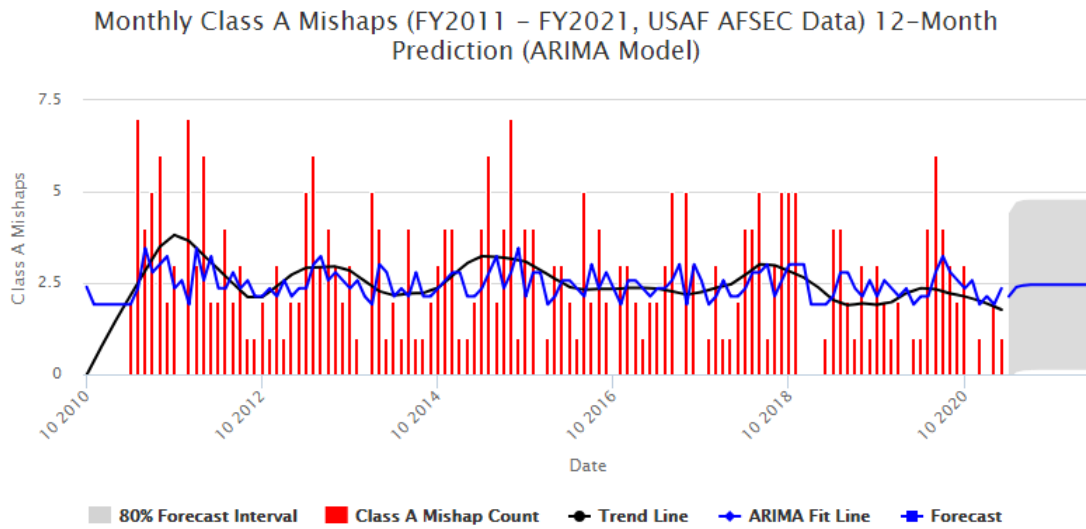


Figure 3.47. AFSAS Class A Seasonal ARIMA Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Results of the Class B Seasonal ARIMA prediction intervals are shown in Figure 3.48.

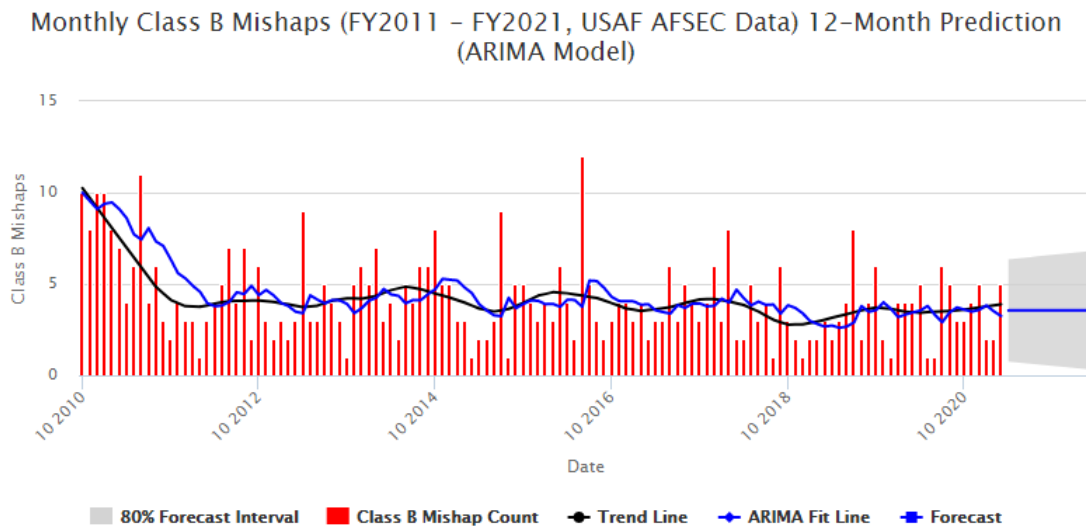


Figure 3.48. AFSAS Class B Seasonal ARIMA Model Prediction with 80% Prediction Intervals Plotted over Monthly Counts

Examination of the Class A and B 1-year Seasonal ARIMA forecasts reveal wide 80%

Confidence Intervals. However, the confidence interval uncertainty did not increase in the Class A model and in the Class B model the confidence interval uncertainty was smooth and grew slowly. These results were unlike what was observed in POC2 (See Section 3.3.5). The Confidence Intervals produced by the Seasonal ARIMA model are still narrower than those generated by the Seasonal Decomposition, however; a result of the Moving Average that is a component of the ARIMA model. The narrower prediction intervals with the Seasonal ARIMA model confirmed that this model will be extremely useful for aviation safety professionals as discussed in Section 3.3.5. See Section 3.5.2 for further details on the use case for the ARIMA model.

3.5 Final Model Selection and Use Cases

3.5.1 Model Selection

The application of POC2 to the AFSAS data confirmed that fitted Seasonal Decomposition and Seasonal ARIMA models would be the best TSA models for predicting USAF monthly mishap incidence. `pmdarima auto.arima()` selected the following optimal parameters for the AFSAS data:

- Class A: Seasonal ARIMA(1, 0, 0)
- Class B: Seasonal ARIMA(1, 0, 1)

3.5.2 Identified Use Cases for Selected Models

AFI 91-202, Paragraph 5.4, requires that USAF commanders and chiefs of safety utilize “historical look-back” approaches complimented by “a proactive, forward looking mishap prevention plan” to “take risk mitigation actions based on sound mishap analysis” (United States Air Force Safety Center 2020, p. 77). The development of the predictive models in this research are designed to assist the annual analysis required by AFI 91-202 to identify, project, and reverse adverse trends. These models also provide MOPs that can assist USAF commanders and chiefs of safety in the identification of “successes, problem areas, and trends to measure safety program effectiveness and guide prevention actions” as required by Paragraph 5.4 (United States Air Force Safety Center 2020, p. 77). In particular, the use of these models assists and compliments the “Study Approach” detailed in Paragraph 5.5.2

of the AFI. The use cases for each model were framed by these requirements.

The Seasonal Decomposition model will assist commanders and chiefs of safety in the identification of historical yearly seasonal trends. These trends identify historical peaks and lows of mishap incidence. Given that the predictions generated by this model for the next FY would incorporate this historical seasonality, safety professionals would subsequently use the predictions of this model to determine the forecasted peak and low months of the coming FY and design the annual safety plan to proactively target the peak month with the intent to reduce the likeliness of a mishap occurring from the top three causes of accidents. Safety professionals at echelon already know the leading causes of both Class A and B mishaps through use of the HFACS.

The Seasonal ARIMA model will assist commanders and chiefs of safety to identify their mishap rate “Band of Normal.” The “Band of Normal” establishes the band of what can be considered the normal or “natural” rate of mishap occurrence for each mishap class. The upper and lower limits of the fluctuations in the fitted model establish the upper and lower limits of this band. Observations of the trend fluctuating up and down within the upper and lower limits established by the fitted ARIMA model in Figure 3.47 supports the idea that the ARIMA model is representative of this concept. Predictions performed by the ARIMA model inform USAF commanders and chiefs of safety if the incidence is predicted to stay within, exceed, or fall below the natural rate of occurrence. The “Band of Normal” also provides a valuable MOP for annual safety program performance:

- If the actual occurrence of mishaps trend above this band, commanders and chiefs of safety should analyze their mishaps for any emerging causal trends for the increase in an effort to curb any additional increases in mishap rates.
- If the actual occurrence of mishaps trend below this band, commanders and chiefs of safety should analyze their safety program for what it was doing to reduce the rate of mishaps.
- If the actual occurrence of mishaps trend within the band, commanders and chiefs of safety should recognize that the rate of mishaps observed occurred within the natural range of occurrence. However, this should not discourage commanders and chiefs of safety from pursuing the reduction of their mishap rates to the lower end of the band through proactive programs.

This use of the “Band of Normal” to detect whether mishap trends are exceeding normal rates is similar to the methodology utilized by International Civil Aviation Organization (ICAO) to detect mishap rate deviation (International Civil Aviation Organization 2022).

It must be noted that the ARIMA use case acknowledges an uncomfortable truth. A goal of zero mishaps was set by Gen (Ret) John P. Jumper, then USAF Chief of Staff, in December 2002 (Kreisher 2006). In subsequent years, the goal of zero mishaps has grown to become a mindset that permeates the USAF Safety Program. However, the ARIMA model use case recognizes that there is a rate of natural mishap occurrence which results from the dangerous and random nature of military aviation coupled with human performance limitations that cannot be fully eliminated. We will discuss this further in Section 4.1.6.

3.6 Normalization of Incidence Prediction

Normalization of incidence counts by frequency is an important methodology in data analytics. “The goal of normalization is to change the values of numeric data in a data set to a common scale, without distorting the differences in the range of values,” such that large values don’t influence the results of further analysis (Jaitley 2018). Therefore, TSA counts should be normalized wherever possible.

The USAF measures aviation incidents in a normalized form as the number of mishaps per 100,000 flying hours. AFI 91-202 states that “standardized mishap metrics... are an effective way to compare the actions and accomplishments of [a] unit” with regard to safety programs. An added advantage of mishap rates standardized on total flying hours (TFH) is that it removes the amount of flying being performed by the USAF as a contributing factor for the Seasonality and “Band of Normal.” Paragraph 5.12.1 identifies the Total USAF Aviation Class A/B Metric as “the number of USAF mishaps (to include flight, flight-related, and aircraft ground operations) and mishap rates per 100,000 flying hours” (United States Air Force Safety Center 2020, p. 80). This results in the following equation:

$$r_{mishaps} = \frac{n_{mishaps} * 100,000}{n_{TFH}}$$

where:

- $r_{mishaps} \in [0, \infty)$ = rate of mishaps per 100,000 TFH for the time period being examined
- $n_{mishaps} \in \mathbb{Z}^+$ = total number of Class A/B mishaps for the time period being examined
- $n_{TFH} \in \mathbb{Z}^+$ = total number of flying hours for the time period being examined.

For the research team's final findings, the normalization equation was applied to the initial findings briefed to AFSEC. The models were also redesigned to predict the rate of mishap occurrence per month, instead of predicting the raw number of mishaps per month, by utilizing monthly TFH data tracked by AFSEC.

3.7 The Aviation Mishap Incidence Predictor Tool

The Aviation Mishap Incidence Predictor (AMIP) is a tool designed to be utilized by AFSEC and USAF safety personnel at echelon to perform analysis and predictive modeling in accordance with AFI 91-202, Paragraph 5.4, requirements. Written in Python and capable of deployment on cloud environments, AMIP was designed to utilize standard data pulls from AFSAS and AFSEC to forecast the next 12 months of mishap incidence. AMIP can also perform analysis on a variety of aviation mishap phenomenon, such as engine mishaps or MDS-specific mishaps, by pulling data from AFSAS for that specific phenomenon. AMIP seamlessly integrates into, assists, and informs the Mishap Prevention Analysis processes specified in AFI 91-202, Paragraph 5.5. See Sections 3.7.11 and 3.7.10 for detailed discussion about use of AMIP in Mishap Prevention Analysis. Source code for AMIP is available by request from Professor Ruriko Yoshida.

3.7.1 Use of AFSAS and AFSEC Standardized Data and Required User Inputs

AMIP was designed to utilize standardized data from AFSEC processes while providing input flexibility to the analyst or safety professional. As long as the input data contains the columns below with the labels and contents specified, AMIP will produce fitted Class A and B models with 12-month forecasts.

AMIP was designed to utilize standard AFSAS data pulls in Excel workbook (.xlsx or .xls) file format. AMIP requires the following data columns from AFSAS data:

- **Mishap Class:** Contains an “A” or “B” identifying the class of mishap.
- **Date:** Contains the day, month, and year a mishap occurred. A specific date format is not required, as AMIP automatically detects the format.

For AMIP to process the mishap data, both of these data columns must be present and spelled exactly as above (including spaces). The order of the columns does not matter. AMIP will disregard all other data columns in the AFSAS Excel file, as no additional data is required to fit the models. Users must specify the worksheet name that contains the mishap data and which row contains the data column labels. “DET Results” is set as the default worksheet name, which is the name automatically assigned by AFSAS to the worksheet containing mishap data. The worksheet row that contains the column labels can change depending on the data pull; the AMIP default is “1” for the first (top) row. For example, if the column labels appear in the third row, a “3” must be input by the user.

There seems to be no AFSEC-standardized data set for TFH; the research team received data sets of different structures during its analysis. AMIP was designed to utilize TFH data in Excel workbook (.xlsx or .xls) file format. AMIP requires the following TFH data columns:

- **Total Flight Hours:** Contains the monthly total number of flight hours that matches the mishap data set. AMIP will convert this to an `integer (int)` or `float` data type.
- **Month:** Contains the single- or double-digit number for a month. AMIP will convert this to an `integer (int)` data type.
- **Month Name:** Contains the character string name of a month. AMIP will keep this as a `string (str)` data type.
- **Year:** Contains the Fiscal Year for the TFH data. AMIP automatically converts this data from FY to CY so that it can match the TFH data to the mishap data.

For AMIP to process the TFH data, all four data columns must be present and spelled exactly as above (including spaces). The order of the columns does not matter, and AMIP will not load any additional data from the Excel file. Users must specify which row contains the data column labels. The worksheet row that contains the column labels can change depending on the data pull; the AMIP default is “1” for the first (top) row. For example, if the column labels appear in the third row, a “3” must be input by the user. AMIP assumes that all TFH

data appears in the first worksheet of the Excel workbook file.

The AMIP user input window is shown in Figure 3.49. In addition to the data paths and column label rows discussed above, the user must specify the data period start and end year and month. Specifically, this is the period the user wishes to model that both data sets cover. Finally, the user must specify the amount of error to be captured by the prediction intervals, known as the α value, which must be entered as a decimal percent. For example, the error value must be set as “0.10” if the user desires 90% Prediction Intervals. The AMIP default value is “0.20” to produce 80% Prediction Intervals.

Aviation Mishap Incidence Predictor (AMIP)




Welcome to the Aviation Mishap Incidence Predictor (AMIP)!

INPUT PARAMETERS

MISHAP DATA

Mishap Data Excel File: D:/Work/NPS/Thesis/AFSEC_Data/AF CLASS A & B,

Mishap Data Worksheet Name: DET Results Format: DET Results

Mishap Data Header Row Number: 3 Format: 1

TOTAL FLYING HOURS (TFH) DATA

Total Flying Hours Data Excel File: D:/Work/NPS/Thesis/AFSEC_Data/flight_hours_by_mo

Total Flying Hours Data Header Row Number: 1 Format: 1

DATA PERIOD

Starting Year of Data Period: 2010 Format: YYYY

Starting Month of Data Period: 10 Format: M or MM

Ending Year of Data Period: 2021 Format: YYYY

Ending Month of Data Period: 3 Format: M or MM

Forecast Confidence Interval Percent Error (Alpha Value): 0.20 Format: 0.20

USER NOTICES

WARNING: Data Period Start and End MUST be the same for both data files!

WARNING: Mishap Data MUST include column names "Mishap Class" and "Date" formatted as shown (including spaces)!

WARNING: TFH Data MUST include the following column names formatted as shown (including spaces):

"Total Flight Hours" "Month" "Month Name" "Year"

NOTE: AMIP assumes monthly data with a 12-month period!

Developed by CPT Ryan Herrmann, MP/49, USA ryan.c.herrmann.mil@army.mil ryan.c.herrmann@gmail.com

Figure 3.49. AMIP User Input Window

To fit models to the data and display the AMIP user dashboard, the user must select the “Run AMIP” button located at the bottom of the input window. To exit and terminate the program, the user must select the “Exit” button. The “Exit” button also terminates hosting of the user dashboard if it has been generated.

If the user wishes to generate models for a different data set, the user must restart AMIP.

3.7.2 Key Non-standard Python Packages

AMIP requires the following non-standard Python packages:

- **pmdarima**: Designed to mimic the format, methodology, and reports of R's implementation of `auto.arima()` in the Python environment. AMIP utilizes the `pmdarima auto.arima()` function to automatically determine the required ARIMA model parameters.
- **statsmodels**: Provides the capability to perform Seasonal Decomposition of time-series data.
- **plotly**: Displays interactive graphics of the fitted models and forecasts on the AMIP user dashboard.
- **dash** and **dash_bootstrap_components**: Construct and host the AMIP user dashboard, which is hosted on the locally on the user's computer and can be accessed via a web browser.
- **pysimplegui**: Displays the AMIP user input window. The package also introduces program multithreading: The user input window is run on the primary thread (Thread 0) while model generation and dashboard hosting is run on a second thread (Thread 1).

3.7.3 Reading In and Cleaning the Data

AMIP reads in the mishap data into a data frame using the Pandas `read_excel()` function. The function only reads in the data columns specified above, and automatically converts the Date data into a datetime data type. The `Mishap Class` column label is then renamed to `Mishap_Class` to remove issues with having spaces in column labels.

AMIP reads in the TFH data into a data frame using the Pandas `read_excel()` function. The function only reads in the data columns specified above. The `Total Flight Hours` and `Month Name` column labels are then renamed to `Total_Hours` and `Month_String` to remove issues with having spaces in column labels.

The datetime formats differ for each data set, and must be aligned for the data sets to be merged. Conversion of the AFSEC TFH year data from FY to CY format was required, as the AFSAS mishap year data is in CY format. This conversion was performed by subtracting 1 from all year values corresponding to an October, November, and December month. A

new `Date` data column was created in the TFH data set by combining the `Year` and `Month` columns. Since `datetime` formats require day information, the day of this new data column was set as the first day of each month. Since the monthly mishap counts automatically format to the last day of each month, the difference between the first and last days of each month was added to the TFH `Date` data.

To remove redundant data from the TFH data frame, the `Month_String`, `Month`, and `Year` columns were dropped, as they were no longer required.

3.7.4 Obtaining Monthly Raw Mishap Counts

The monthly mishap counts were obtained using the Pandas `pivot_table()` function, where the index was set as `Date` grouped by month and the columns by `Mishap_Class`. The `size` aggregating function provided the counts. This resulted in a data frame containing `A`, `B`, and `N` data columns indexed by the end day of each month. The `N` column contained the sum of Class `A` and `B` mishaps for a given month. The column labels were renamed to `A_Raw`, `B_Raw`, and `N_Raw` respectively to eliminate confusion with the raw mishap counts and the mishap rates to be calculated.

3.7.5 Handling Missing Months in the Time Series

Before merging the mishap and TFH data frames, it was necessary to ensure all months were present in the data between the start and end periods to satisfy TSA modeling requirements. The Pandas `date_range().difference()` function provided a list of any months missing from the mishap counts data frame. If this list contains any months, three corresponding lists are built containing the number of zeros required for Class `A`, `B`, and `N` mishaps. This data is then inserted into a data frame, which is concatenated to the bottom of the mishap counts data frame. This amended mishap counts data frame is then sorted according to `Date` to ensure the missing months are in their proper place in the time-series.

3.7.6 Merging the Data Sets and Calculating Mishap Rate

The TFH data is then added to the mishap counts data via left inner join onto the mishap counts data frame by `Date`. Performing the merge after the check for missing months ensures

that any months where zero mishaps occurred still has its corresponding hours data for the mishap rate calculations.

AMIP now normalizes the mishap counts into mishap rates using the equation specified by AFI 91-202 seen in Section 3.6. This results in new A, B, and N mishap rate data columns. The data is ready for model fitting and forecasting.

3.7.7 Modeling and Forecasting

Seasonal Decomposition of the time-series data was performed for each class separately using the `seasonal_decompose()` function from the `statsmodels.tsa.seasonal` subpackage of the `statsmodels` package. This resulted in two data frames containing `seasonal`, `trend`, `residuals`, `seasonal_adj`, and `actuals` data columns.

Fitting of ARIMA models for p , d , and q parameters was performed using the `auto.arima()` function from the `pmdarima` package. Two models were fit for each mishap class: one that does not account for seasonality and one that does. The `predict_in_sample()` function was then used to determine the monthly fitted values for each model, which were combined into a single data frame.

Prior to model forecasting, a data frame of naive forecasts was constructed where the forecast for time step t is simply the value at time step $t - 1$ of the original data.

One-step-ahead forecasts were then constructed for the Seasonal Decomposition and ARIMA models of each mishap class. To evaluate model performance, the MSE and SMAPE values for the forecasts were calculated by comparing the one-step-ahead forecasts to the naive forecasts. The performance evaluation values for each model were captured in a single data frame for later display. AMIP will automatically select which ARIMA model to utilize based off these calculations. The default is the Seasonal ARIMA model; AMIP will only select the standard ARIMA model if the model's SMAPE is less than the SMAPE of the Seasonal ARIMA model.

3.7.8 AMIP Dashboard

AMIP displays model fits and forecasts in an interactive user dashboard hosted locally on the user's computer by the `dash` package. The fitted models are displayed on top of the historical

monthly mishap rates. The forecasts are displayed as hashed lines and contain confidence intervals to the amount of error (α value) specified by the user. The user dashboard has four pages:

- Main Dashboard
- Class A Dashboard
- Class B Dashboard
- Model Performance and Fit Dashboard

The user can navigate between the pages using the “Dashboards” drop-down menu in the upper right of the each page. Each page also displays the mishap normalization rate utilized, the path for the input mishap data file, and the path for the input TFH data file at the bottom.

The display of each model is interactive. Users can use the mouse to hover over data, displaying the values for each time step (Figure 3.50). All data displayed can be toggled on and off by clicking the legend item for the particular data. Data not being displayed is “faded out” in the legend. Users can inspect pre-configured subsets of the data using the buttons located at the top left of each graph:

- 1M: Displays the last month of data.
- 3M: Displays the last three months of data.
- 6M: Displays the last six months of data.
- 1Y: Displays the last year of data.
- 3Y: Displays the last three years of data.
- YTD: Displays the last year of data from today’s date.
- ALL DATA: Displays the entire period of data available.

The interactive displays also include utility buttons in the upper right corner of each graphic (Figure 3.51). These buttons allow the user to take screenshots of the graphic, zoom, select data, and reset the graphic to its original state.

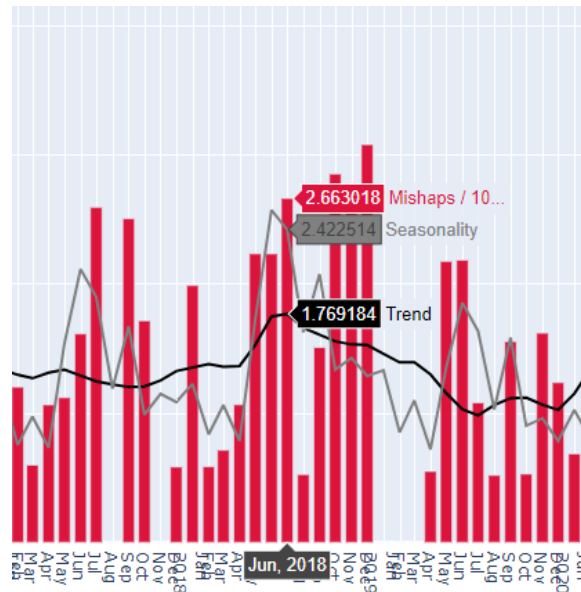


Figure 3.50. AMIP Dashboard Interactive Graphics Mouse Hover Feature. Hovering the mouse will display the values for each time step.



Figure 3.51. AMIP Dashboard Interactive Graphics Utility Buttons

The Main Dashboard (Figure 3.52) displays the both Class A and B models simultaneously, and is intended to be used as a “roll-up” dashboard particularly suited for use with senior leaders or for comparing trends between mishap classes.



Figure 3.52. AMIP Main Dashboard. Designed as a “roll-up” dashboard displaying both models that is suited for use with senior leaders or for comparing trends between mishap classes.

The dashboard for each class (Figures 3.53 and 3.54) is designed to allow analysts a more detailed view of the data for that particular mishap class. It is intended to enable detailed analysis of trends.



Figure 3.53. AMIP Class A Dashboard. Designed as a detailed dashboard for analysis of Class A data.

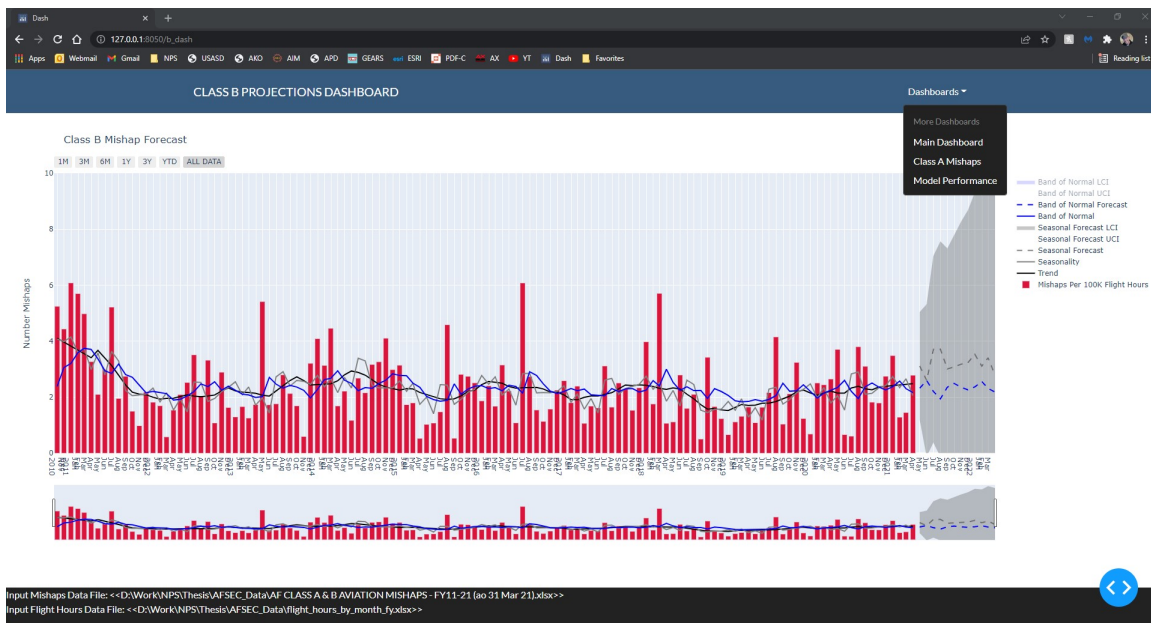


Figure 3.54. AMIP Class B Dashboard. Designed as a detailed dashboard for analysis of Class B data. The dashboard drop-down menu can also be seen in the upper right.

The Model Fit and Performance Dashboard (Figure 3.55) displays the model fit summaries for each ARIMA model and the SMAPE values for all models. It is intended to be used by analysts to understand model fit and performance.

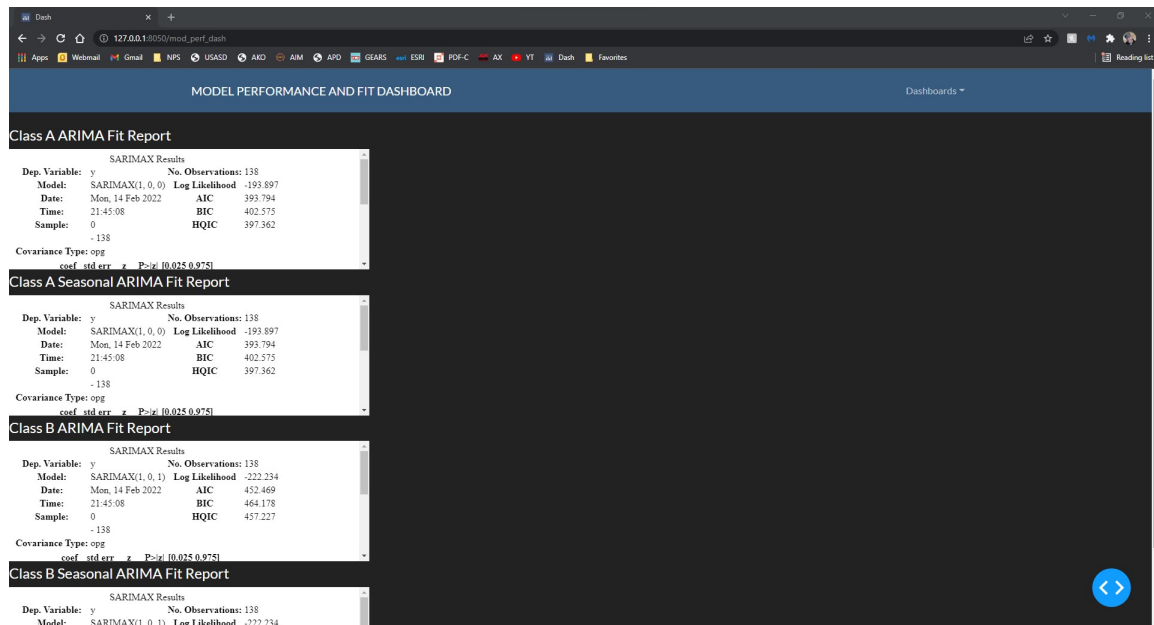


Figure 3.55. AMIP Model Fit and Performance Dashboard. Designed to be used by analysts to evaluate model fit and performance.

The AMIP user dashboard stays active for as long as desired. The user must select the “Exit” button on the user input window and exit the web browser to terminate the user dashboard. If the user wishes to generate models for a different data set, the user must restart AMIP.

3.7.9 Deployment on Cloud Environments

Deployment of AMIP via Docker container ensures all required Python packages accompany the program code. The Docker container can be deployed on a variety of cloud environments, such as the USAF’s CloudOne or Amazon Web Services (AWS). A separate document explaining how to deploy AMIP will accompany the program.

3.7.10 Using AMIP in the “Study Approach” to Mishap Prevention Analysis

The “Study Approach” to Mishap Prevention Analysis is “a detailed examination of a specific problem area through the use of a systematic process... [that] typically follows the format of the Scientific Method” (United States Air Force Safety Center 2020, p. 77–78). AMIP is fully capable of participating in and assisting the “Study Approach,” as it can construct Seasonality and “Band of Normal” models for any phenomenon that is captured by AFSAS data. For example, if a USAF researcher is performing analysis to “determine a trend in frequency of *X* and identify possible explanations of this trend” (United States Air Force Safety Center 2020, p. 78), AMIP can be used to identify the frequency trends for *X* and guide researchers to examine its causes using known HFACS statistics. *X* can be any phenomenon captured by AFSAS data to include mishaps resulting from particular parts failures (engines, landing gear, etc.) and the mishap history of particular aircraft MDS. Once the trends for *X* are identified by AMIP, the researcher can analyze HFACS statistics for potential causes. For example, if AMIP reveals peak seasonality of *X* occurs in May, the researcher can examine the top three HFACS codes of *X* to derive possible causes of that peak seasonality. AMIP supplements and informs the researcher’s analysis in the Study Approach to Mishap Prevention Analysis.

3.7.11 Using AMIP in the “Target Approach” to Mishap Prevention Analysis

The “Target Approach” to Mishap Prevention Analysis is similar in methodology to the “Study Approach,” but develops recommendations and priorities based on the frequency and severity of mishap causes determined by the analysis (United States Air Force Safety Center 2020, p. 77). While AMIP supplements and informs the researcher’s analysis in this approach as it does in Section 3.7.10, it also assists commanders by providing useful MOPs for frequency and severity of mishaps as described in Section 3.5.2. Performing analysis with AMIP will “help prevent mishaps and present conclusions drawn in a useful format that can be applied to prevention programs,” and enable follow-up analysis to determine mishap program effectiveness, as specified by AFI 91-202, Paragraph 5.6 (United States Air Force Safety Center 2020, p. 78).

3.7.12 Validation of AMIP Use in Mishap Prevention Analysis — T-38 MDS Case Study

To validate the capability of AMIP to handle filtered data from AFSAS, the research team conducted analysis of mishap trends for the T-38 MDS. The T-38 mishap history was pulled from AFSAS and the TFH data was received from AFSEC personnel. Figures 3.56 and 3.57 show the resulting Class A and B models for the T-38. The models do not possess high forecast accuracy due to the low number of mishaps in the data set, a combined total of 21 Class A and B mishaps across 10 years of data. However, the models still provide useful insights into the mishap rate Seasonality and “Band of Normal” for the T-38 MDS. Therefore, AMIP is capable of participating in and being a useful tool for both Mishap Prevention Analysis processes.

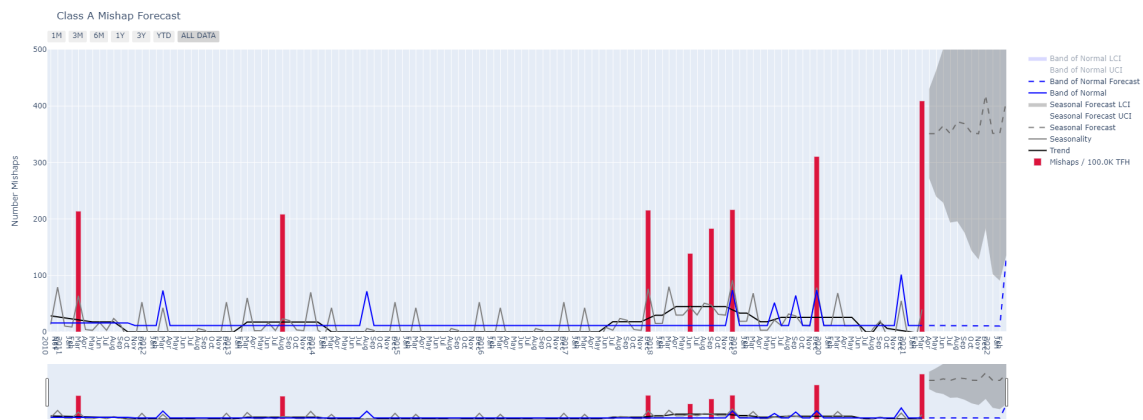


Figure 3.56. AMIP Class A Models for the T-38 MDS. While the forecasts are not necessarily accurate, peak and low Seasonality and the “Band of Normal” for T-38 mishaps are still identifiable.

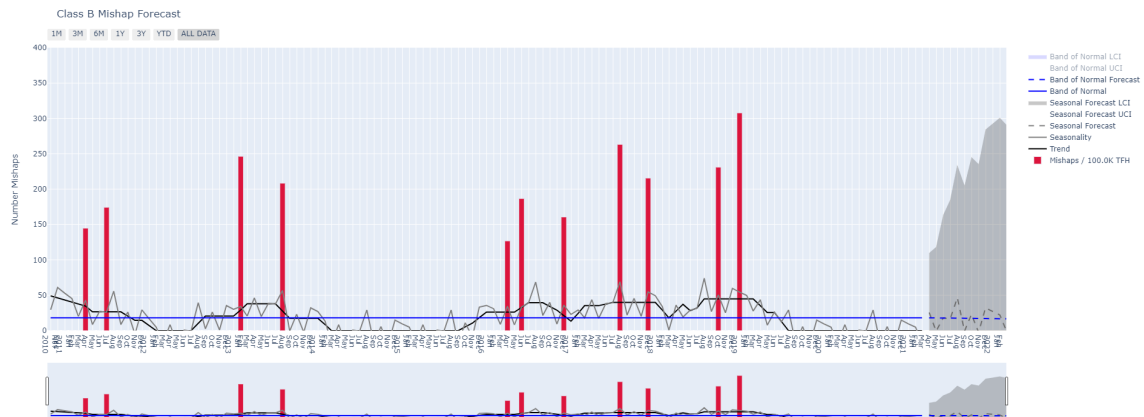


Figure 3.57. AMIP Class B Models for the T-38 MDS. While the forecasts are not necessarily accurate, peak and low Seasonality and the “Band of Normal” for T-38 mishaps are still identifiable.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:

Findings

4.1 Number of Mishaps per Month

The Initial Findings utilized models built to forecast the raw number of Class A and B mishaps per month. Feedback from AFSEC resulted in modifying the models to forecast the Class A and B monthly mishap rate by normalizing the historical data using the formula discussed in Section 3.6. See Section 4.2 for the Final Findings.

4.1.1 Class A Mishap Seasonality

Examination of the Seasonality for Class A mishaps (Figure 4.1) revealed that peak seasonality occurs in May for the ten year duration of AFSAS mishap data. Low seasonality of Class A mishaps occur either in January or March, changing each year with no discernible pattern. These seasonality trends were consistent across the ten years of AFSAS data examined.

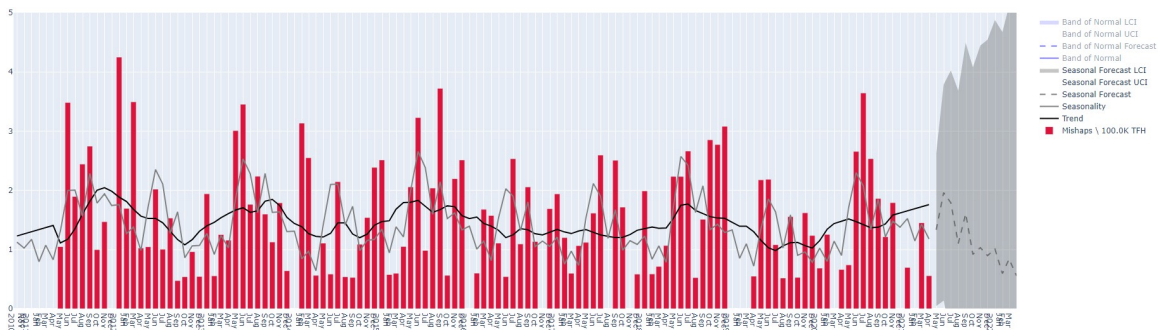


Figure 4.1. AFSAS Class A Mishap Seasonality — Number of Mishaps. Each red bar depicts the historical number of mishaps per month. The solid black line represents the Trend component of the Seasonal Decomposition. The solid grey line depicts the Seasonality component of the Seasonal Decomposition. The hashed grey line represents the Seasonal forecast for the next 12 months, with the shaded area denoting the 80% Prediction Interval. Peak seasonality occurs in May each year; low seasonality occurs in either January or March, depending on the year. Note that these trends are consistent across all ten years of data.

4.1.2 Class A Mishap Band of Normal

Remember that the “Band of Normal,” or the band of natural occurrence, is measured as:

- Upper Bound = Largest Value (Highest Peak) of ARIMA Model Fit
- Lower Bound = Smallest Value (Lowest Valley) of ARIMA Model Fit.

Examination of the “Band of Normal” for Class A mishaps (Figure 4.2) revealed that it is natural to have two or three mishaps occur each month. Note that this band remains steady across all ten years of AFSAS data.

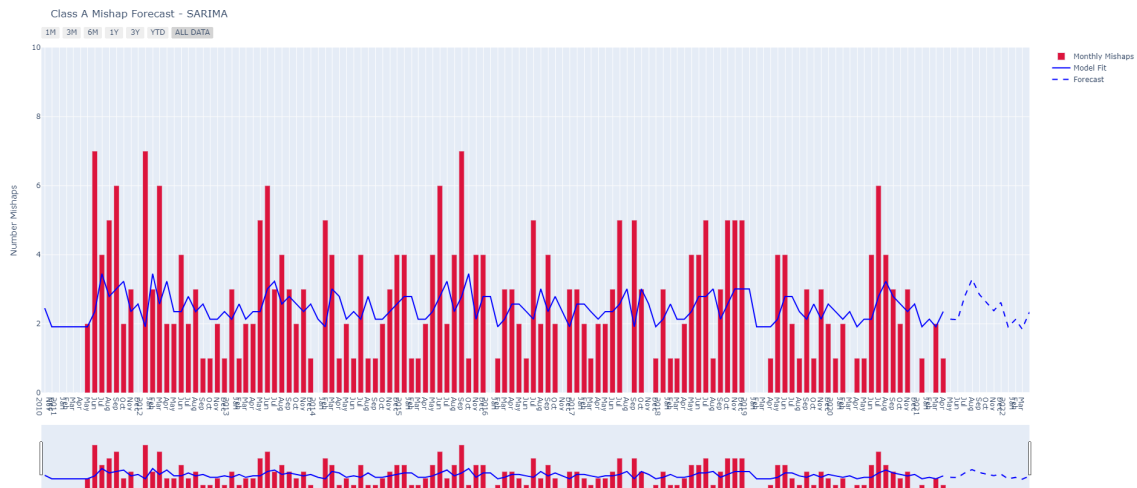


Figure 4.2. AFSAS Class A Mishap “Band of Normal” — Number of Mishaps. Each red bar depicts the historical number of mishaps per month. The solid blue line represents the Seasonal ARIMA model fit. The highest peak and lowest valley of the blue line define the upper and lower bounds of the “Band of Normal.” The hashed blue line represents the Seasonal ARIMA forecast for the next 12 months. The “Band of Normal” for Class A mishaps is 2-3 mishaps per month.

4.1.3 Class B Mishap Seasonality

Examination of the Seasonality for Class B mishaps (Figure 4.3) revealed that mishaps spike in June and July, with the peak seasonality occurring in June and a slight decrease occurring in July. Class B mishap seasonality significantly decreases in August, then slowly

decreases until low seasonality in May. These seasonality trends were consistent across all ten years of AFSAS data.

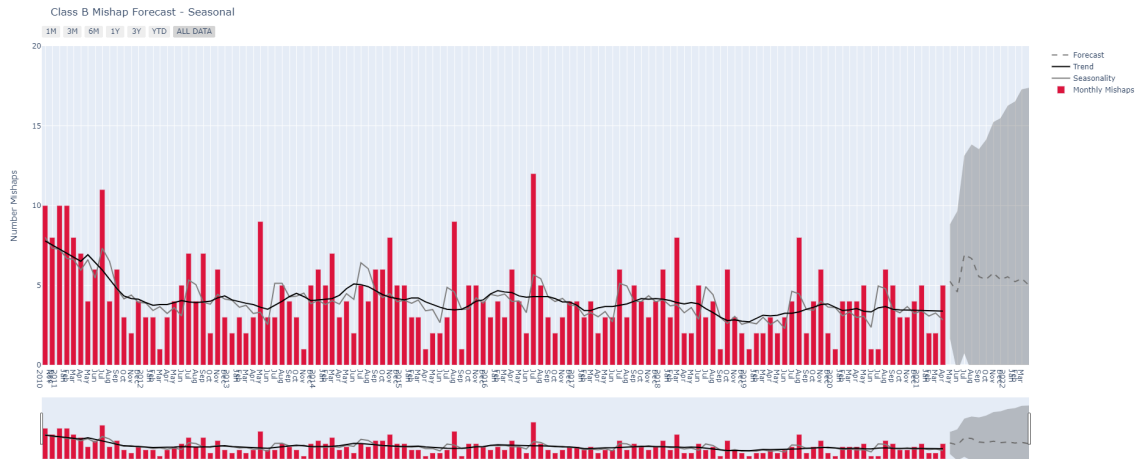


Figure 4.3. AFSAS Class B Mishap Seasonality — Number of Mishaps. Each red bar depicts the historical number of mishaps per month. The solid black line represents the Trend component of the Seasonal Decomposition. The solid grey line depicts the Seasonality component of the Seasonal Decomposition. The hashed grey line represents the Seasonal forecast for the next 12 months, with the shaded area denoting the 80% Prediction Interval. Mishaps spike in June and July each year, then dramatically decrease in August. The number of mishaps per month then slowly decreases until the low seasonality in May. Note that these trends are consistent across all ten years of data.

4.1.4 Class B Mishap Band of Normal

Examination of the Band of Normal for Class B mishaps (Figure 4.4) revealed that it is natural to have three or four mishaps occur each month. Across the ten years of AFSAS data analyzed, the band decreased in steps representing small, incremental decreases in Class B mishap normality followed by lengthy plateaus.

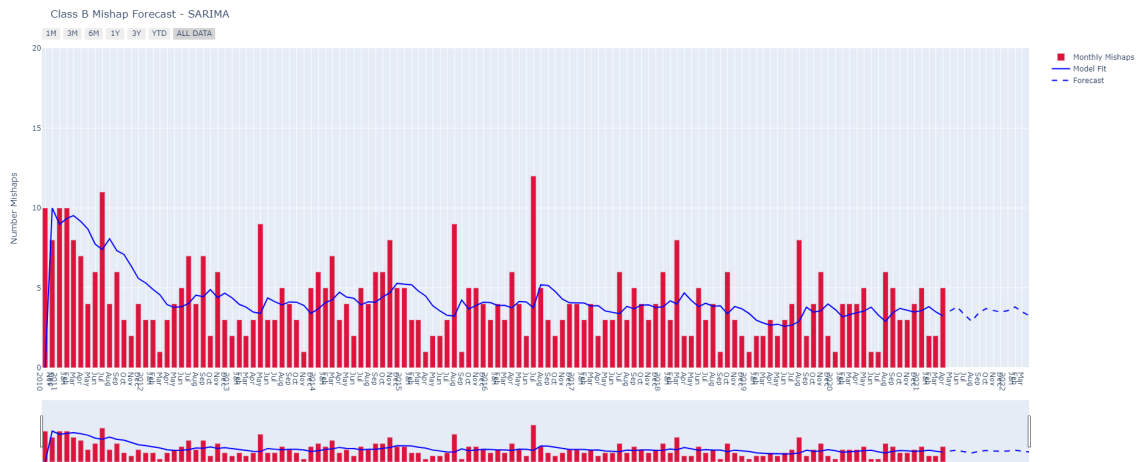


Figure 4.4. AFSAS Class B Mishap “Band of Normal” — Number of Mishaps. Each red bar depicts the historical number of mishaps per month. The solid blue line represents the Seasonal ARIMA model fit. The highest peak and lowest valley of the blue line define the upper and lower bounds of the “Band of Normal.” The hashed blue line represents the Seasonal ARIMA forecast for the next 12 months. The “Band of Normal” for Class B mishaps is 3-4 mishaps per month.

4.1.5 AMIP Model Predictive Power

Due to high, unsatisfactory SMAPE values reported by AMIP, the research team wanted to determine the true predictive power and accuracy of the models. To perform this analysis, the team utilized the forecasts produced by the Class A Seasonal Decomposition model (Figure 4.5) and checked the values against the actual occurrence of Class A mishaps. The count of actual mishaps was extracted from a list of Class A Mishap Details published in the Department of the Air Force (DAF) Mishap Summary produced by AFSEC. At the time of conducting the initial findings, the most complete data available was as of November 2021. The team selected Near, Mid, and Far forecast horizons to analyze predictive performance at different points in the forecast period:

- Near: June 2021;
- Mid: August 2021;
- Far: November 2021.

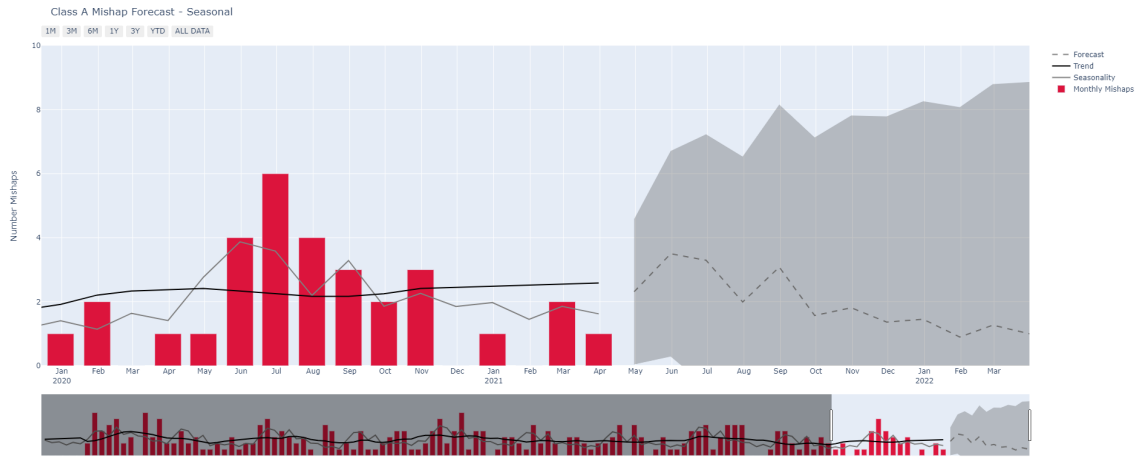


Figure 4.5. AFSAS Class A Mishap Seasonal Model Forecast — Number of Mishaps. The forecast depicted by the hashed grey line was utilized to perform model predictive power analysis as depicted in Table 4.1

Table 4.1. Class A Seasonal Model Predictive Performance

Forecast Horizon	Model Forecast	Actual Mishaps
Near (JUN21)	3.3	3
Mid (AUG21)	3.1	3
Far (NOV21)	1.4	2

Rounding the forecasted number of mishaps to the nearest whole number places the Near and Mid forecast horizons in line with the actual number of mishaps that occurred, indicating strong Near and Mid predictive power. The Far forecast is slightly off, which can be expected given the distant horizon and the growing levels of uncertainty.

Examination of Table 4.1 reveals that the near, mid, and far predictive performance of the Class A Seasonal model is nearly spot on. However, one cannot have 0.3 or 0.1 of a mishap. If USAF analysts rounded the forecast values to the nearest whole number, the Near and Mid predictive performance is accurate; the Far predictive performance is just a bit low.

Given the small values the model is attempting to forecast coupled with the large amount of randomness present in mishap occurrence, this predictive performance is excellent. This also revealed that the large SMAPE values are a result of the small values being predicted by the model.

4.1.6 Natural Mishap Rate and Setting Mishap Reduction Goals

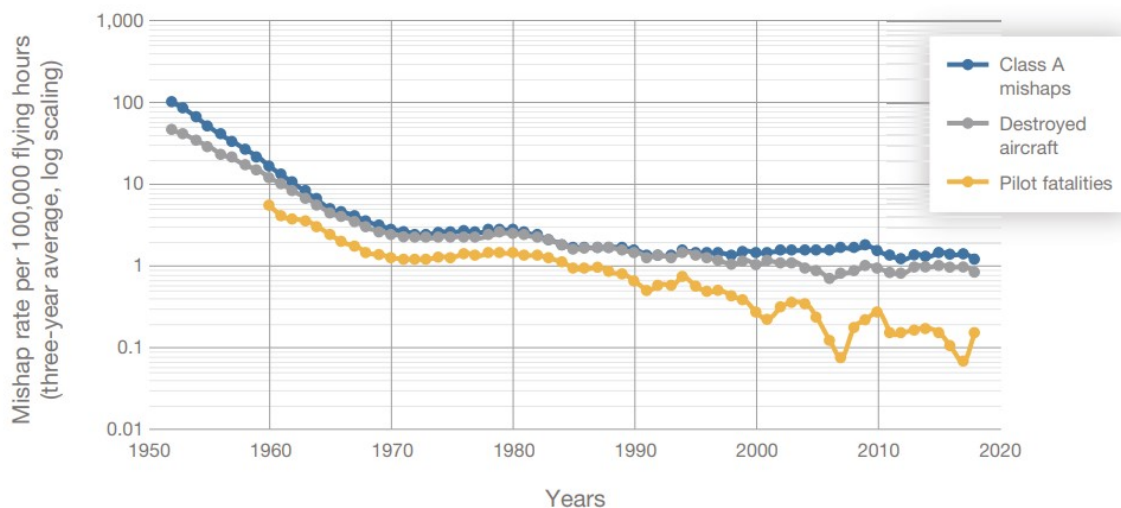
As briefly discussed in Section 3.5.2, the natural mishap rate of occurrence demonstrated by the “Bands of Normal” revealed an uncomfortable truth about mishap prevention: mishaps will occur simply because of the randomness, human performance limitations, and danger inherent to military aviation. This uncomfortable truth directly contradicts the goal set by Gen (Ret) John P. Jumper in December, 2002: that any number of mishaps is unacceptable (Kreisher 2006). The drive “toward zero mishaps” permeates USAF safety culture throughout the force. The fact that the Class A and B mishap rates have remained steady across 10 years of data, largely fluctuating within their “Band of Normal,” strongly suggests three conclusions:

1. The systemic issues of military aviation operations that result in mishap occurrence have largely been mitigated or eliminated;
2. The sufficient mitigation or elimination of systemic issues leaves randomness, limitations of human performance, and characteristics inherent to military aviation (e.g. danger) to cause military aviation mishaps; and
3. The fluctuation of mishap incidence observed in the ten years of AFSAS data is likely to continue, and may be marginally reduced, but cannot be fully eliminated.

A report published by the RAND Corporation in 2019 supports these conclusions. The report, performing statistical analysis on USAF mishap rates from 1950 to 2018, concluded that “major improvements in flight safety have been achieved,” with the rate of aircraft lost being significantly reduced from 23.6 aircraft lost per 100,000 TFH during the 1950s to less than one aircraft lost per 100,000 TFH between 2010 and 2018 (Light et al. 2019, p. 10). Figure 4.6, which is Figure 1 from the RAND report, demonstrates this precipitous reduction. The report also concluded that aircraft MDS introduced more recently “tend to exhibit lower mishap rates” due to integration of safety technology, increased manufacturing and parts reliability, and improved safety practices (Light et al. 2019, p. 11). These improvements

mitigate or eliminate military aviation operational systemic issues that result in mishap occurrence. The decrease demonstrated in Figure 4.6 of aircraft mishap and loss rates can be directly tied to these improvements. Additionally, the slowing of reductions to mishap and aircraft loss rates observed in Figure 4.6 indicates that military aviation safety is now attempting to reduce mishap incidence resulting from randomness, human performance factors, and the introduction of new technologies.

FIGURE 1
Trends in U.S. Air Force Aircraft Mishap Rates over Time



SOURCE: Authors' analysis of data accessed August 3, 2019, from Air Force Safety Center, undated.
NOTE: Each data point represents a three-year trailing average.

Figure 4.6. Historical Trend in USAF Aviation Mishap Rates between 1950 and 2018. The slowing of reductions to mishap and aircraft loss rates indicates that military aviation safety is now attempting to reduce an already incredibly low rate of mishap incidence. The sufficient mitigation or elimination of systemic issues across this period leaves randomness, limitations of human performance, and characteristics inherent to military aviation to cause military aviation mishaps, which cannot be eliminated. Note that the mishap rate y-axis is a logarithmic scale. Source: Light et al. (2019, p. 5), Figure 1.

While reduction of mishaps to the lowest rate possible is an ideal goal for any safety program, it must be acknowledged that achieving the goal of a zero mishap rate is not realistic for military aviation. Given the amount of danger and randomness inherent in military aviation

operations combined with the limitations of human performance, realistic goals for the reduction of mishap rates must be set using measurable criteria and analysis of historical trends. The model use cases established in Section 3.5.2 provide the MOPs required to set realistic goals for military aviation safety programs and can assist in the targeting of human performance factors for mitigation.

4.2 Mishap Rate per Total Flying Hours

Normalization of the data to mishap rate per 100,000 TFH did not change the trends identified in the Initial Findings. However, normalization did result in the identification of three important additional conclusions:

1. The trends identified in the Initial Findings are not a result of TFH frequency;
2. The peak seasonal incidence of Class A mishaps exceeds the “Band of Normal;” and
3. The seasonal incidence of Class B mishaps is within the “Band of Normal.”

It is important for USAF safety professionals to identify why peak seasonal incidence occurs in May and June each year for Class A and B mishaps respectively, since the timing of these peaks is no longer attributable to any increase in TFH.

4.2.1 Class A Mishap Seasonality

Examination of the Seasonality for Class A mishaps (Figure 4.7) revealed that peak seasonality occurs in May for the ten year duration of AFSAS mishap data. In May, 2020, this peak was 2.9 Class A mishaps per 100,000 TFH. Low seasonality of Class A mishaps occurs either in January or March, changing each year with no discernible pattern. The low was 1.14 Class A mishaps per 100,000 TFH in January, 2021, and 1.17 Class A mishaps per 100,000 TFH in March, 2021. These seasonality trends were consistent across the ten years of AFSAS data examined.

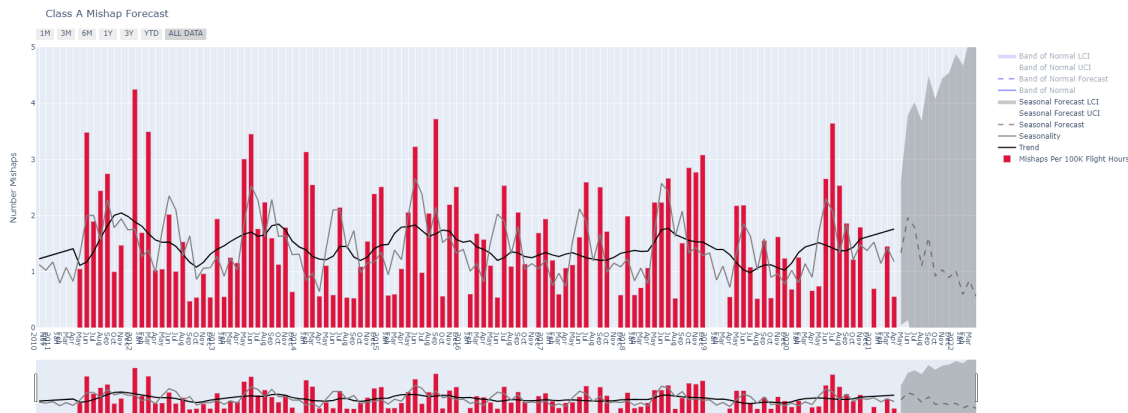


Figure 4.7. AFSAS Class A Mishap Seasonality — Mishap Rate. Each red bar depicts the historical number of mishaps per month. The solid black line represents the Trend component of the Seasonal Decomposition. The solid grey line depicts the Seasonality component of the Seasonal Decomposition. The hashed grey line represents the Seasonal forecast for the next 12 months, with the shaded area denoting the 80% Prediction Interval. Peak seasonality occurs in May each year; this value was 2.29 mishaps per 100,000 TFH in 2020. Low seasonality occurs in either January or March, depending on the year; these values were 1.14 mishaps per 100,000 TFH in January, 2021, and 1.17 mishaps per 100,000 TFH in March, 2021. Note that these trends did not change from the Initial Findings.

4.2.2 Class A Mishap Band of Normal

Examination of the “Band of Normal” for Class A mishaps (Figure 4.8) revealed that the natural Class A mishap rate is 1.1 - 1.9 mishaps per 100,000 TFH per month. Note that this band remains steady across all ten years of AFSAS data.

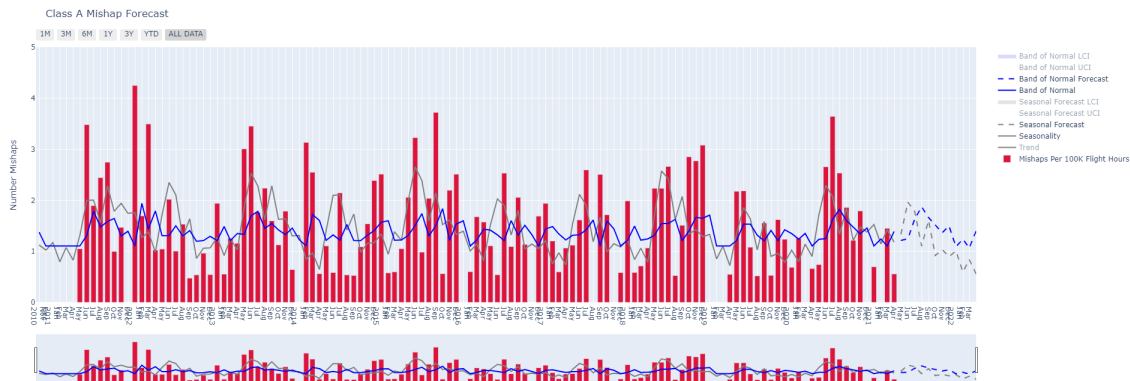


Figure 4.8. AFSAS Class A Mishap “Band of Normal” — Mishap Rate. Each red bar depicts the historical number of mishaps per month. The solid blue line represents the Seasonal ARIMA model fit. The highest peak and lowest valley of the blue line define the upper and lower bounds of the “Band of Normal.” The hashed blue line represents the Seasonal ARIMA forecast for the next 12 months. The “Band of Normal” for Class A mishaps is 1.1 - 1.9 mishaps per 100,000 TFH per month.

4.2.3 Class B Mishap Seasonality

Examination of the Seasonality for Class B mishaps (Figure 4.9) revealed that mishaps spike in June and July, with the peak seasonality occurring in June and a slight decrease occurring in July. In June, 2020, this peak was 2.83 Class B mishaps per 100,000 TFH. Class B mishap seasonality significantly decreases in August, then slowly decreases until low seasonality in May. In May, 2021, this low was 2.06 Class B mishaps per 100,000 TFH. These seasonality trends were consistent across the ten years of AFSAS data examined.

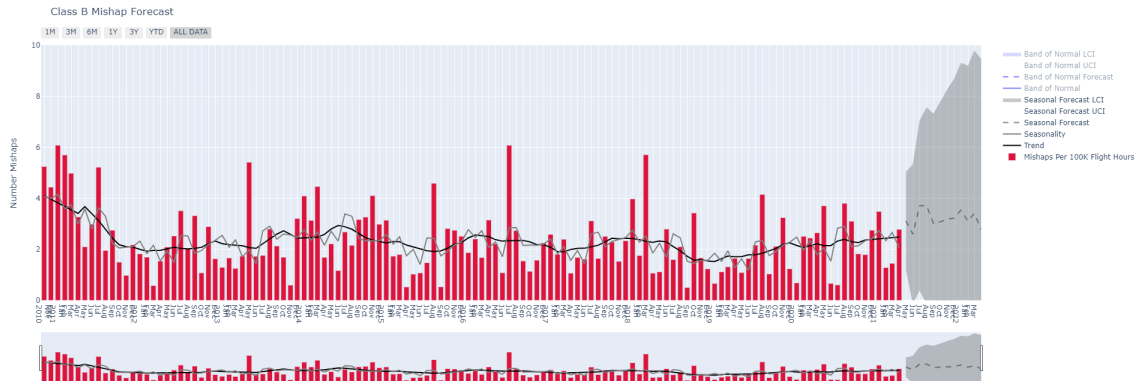


Figure 4.9. AFSAS Class B Mishap Seasonality — Mishap Rate. Each red bar depicts the historical number of mishaps per month. The solid black line represents the Trend component of the Seasonal Decomposition. The solid grey line depicts the Seasonality component of the Seasonal Decomposition. The hashed grey line represents the Seasonal forecast for the next 12 months, with the shaded area denoting the 80% Prediction Interval. Peak seasonality occurs in June each year; this value was 2.83 mishaps per 100,000 TFH in 2020. Low seasonality occurs in May; this value was 2.06 mishaps per 100,000 TFH in 2021. Note that these trends did not change from the Initial Findings.

4.2.4 Class B Mishap Band of Normal

Examination of the “Band of Normal” for Class B mishaps (Figure 4.10) revealed that it is natural Class B mishap rate is 1.85 - 2.99 mishaps per 100,000 TFH per month. Across the ten years of AFSAS data analyzed, the band decreased in steps representing small, incremental decreases in Class B mishap normality followed by lengthy plateaus.

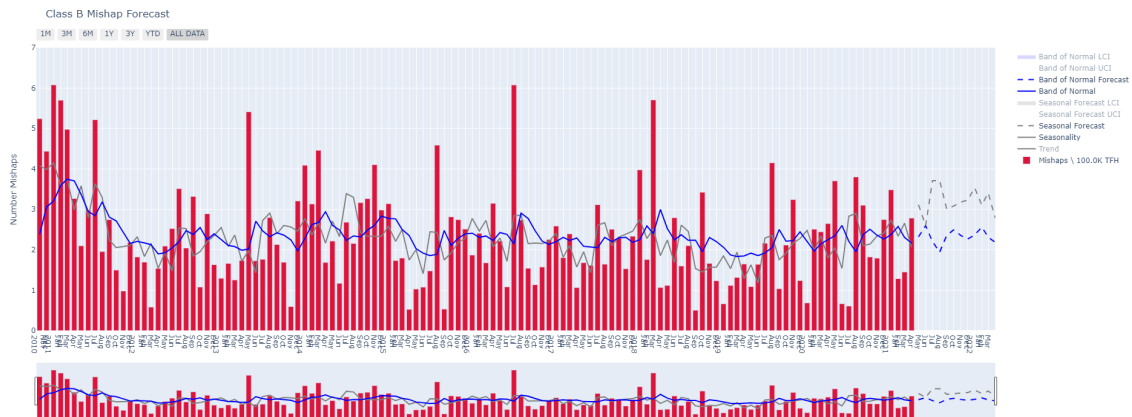


Figure 4.10. AFSAS Class B Mishap “Band of Normal” — Mishap Rate. Each red bar depicts the historical number of mishaps per month. The solid blue line represents the Seasonal ARIMA model fit. The highest peak and lowest valley of the blue line define the upper and lower bounds of the “Band of Normal.” The hashed blue line represents the Seasonal ARIMA forecast for the next 12 months. The “Band of Normal” for Class B mishaps is 1.85 - 2.99 mishaps per 100,000 TFH per month.

4.2.5 Comparing Seasonality to the Band of Normal

Tables 4.2 and 4.3 compare the most recent peak and low seasonal incidence values to the “Band of Normal” for each class of mishaps. Examination of Table 4.2 revealed that the most recent peak rate of Class A mishaps exceeds the “Band of Normal.” Examination of Table 4.3 revealed that the most recent year of seasonal incidence for Class B mishaps is within the “Band of Normal.”

Table 4.2. Comparison of Class A Seasonality to “Band of Normal”

Class A		
Model	Low	High
Seasonality	1.14 (JAN21)	2.29 (MAY20)
Band of Normal	1.10	1.90

Examination of table revealed that the most recent peak rate of Class A mishaps exceeds the “Band of Normal.” The fact that the seasonal mishap rates are mostly within the “Band of Normal” reinforces the idea that mishap rates are settling to their natural rates of occurrence. Section 4.1.6 discusses the significance of this observation.

Table 4.3. Comparison of Class B Seasonality to “Band of Normal”

Class B		
Model	Low	High
Seasonality	2.06 (MAY20)	2.83 (JUN21)
Band of Normal	1.85	2.99

Examination of table revealed that the most recent year of seasonal incidence for Class B mishaps is within the “Band of Normal.” The fact that the seasonal mishap rates are within the “Band of Normal” reinforces the idea that mishap rates are settling to their natural rates of occurrence. Section 4.1.6 discusses the significance of this observation.

The fact that the seasonal mishap rates are mostly within the “Band of Normal” reinforces the idea that mishaps rate are settling to their natural rates of occurrence, as discussed in Sections 3.5.2 and 4.1.6.

4.2.6 AMIP Model Predictive Power

AFSEC feedback from the initial findings requested an examination of AMIP forecast performance across the entire forecast period. The team utilized the forecasts produced by both Class A models (Figure 4.11) and compared the values to the actual Class A mishap rate per 100,000 TFH. The count of actual mishaps was extracted from a list of Class A Mishap Details published in the DAF Mishap Summary produced by AFSEC. At the time of conducting the final findings, all data was available to analyze the complete forecast period. These values are tabulated in Table 4.4.

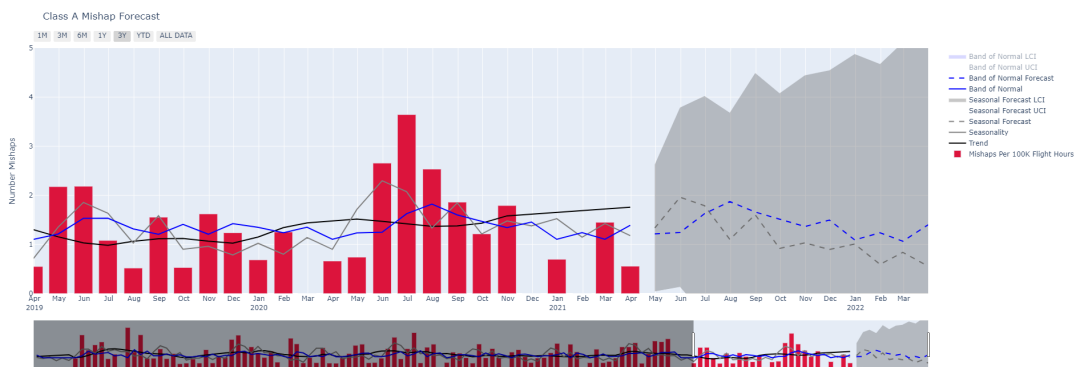


Figure 4.11. AFSAS Class A Mishap Model Forecasts. The forecasts depicted by the hashed grey and blue lines were utilized to perform model predictive power analysis as depicted in Table 4.4

Table 4.4. Class A Model Predictive Performance — Mishap Rate

Forecast Horizon	Seasonal Forecast	ARIMA Forecast	Actual Mishap Rate
APR21	1.335	1.214	1.130
MAY21	1.962	1.245	3.017
JUN21	1.790	1.627	1.789
JUL21	1.103	1.873	0.000
AUG21	1.606	1.661	1.781
SEP21	0.919	1.513	0.675
OCT21	1.032	1.365	0.000
NOV21	0.897	1.497	0.854
DEC21	1.009	1.092	0.976
JAN22	0.595	1.239	0.894
FEB22	0.839	1.065	0.875
MAR22	0.557	1.401	0.659

Examination of the table revealed that the predictive performance of both Class A models is moderate. With modification of the models to predict mishap rate, the values of the forecasted rates are closer to zero, increasing forecasting difficulty. The models still provide useful MOPs for commanders and chiefs of safety with this predictive performance.

Examination of Table 4.4 revealed that the predictive performance of both Class A models is moderate. When considering the amount of randomness inherent in the occurrence of a Class A mishap, the predictions are not too bad. With modification of the models to predict mishap rate, the values of the forecasted rates are closer to zero than when the models were forecasting the raw number of mishaps. This caused SMAPE values to increase significantly, indicating the inherent difficulty of forecasting these values. However, the research team still believes the MOPs provided by each of the models are of value to USAF commanders and chiefs of safety.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:

Conclusion

5.1 Enumerated Answers to Research Questions

Over a year of research effort resulted in the following conclusions:

1. It is possible to predict the number of military aviation Class A and B mishaps using historical data from existing safety data systems. Safety system data does not contain enough predictors for SLR models. TSA models provide a level of predictive capability that provide MOPs to USAF Commanders and Chiefs of Safety. These MOPs can be utilized to develop proactive safety programs in accordance with AFI 91-202 Paragraph 5.4 and 5.5 requirements.
2. Seasonal Decomposition and ARIMA TSA models are the best regression methodologies to construct predictive models for military aviation safety incidence. These methodologies are employed in a tool which can be incorporated into USAF safety practices.
3. We were unable to determine the leading causes for Class A and B mishaps and whether those causes are changing across time. USAF PSI protections prevented TSA of HFACS. However, such analysis is technically possible if HFACS incidence is modeled in the same way as Class A and B mishaps in this research. The performance of any statistical analysis on HFACS frequency (e.g. leading HFACS causes for each mishap class) would not be better than analysis currently performed by AFSEC.
4. AFSEC already performs analysis to determine which aircraft types or models have the highest rates of Class A and B mishaps. The research team felt that building this analytical capability into AMIP would stretch the research beyond its timetable to no additional benefit or insight.

5.2 Recommendations for Future Research

The research represents an initial exploration into application of advanced data analytics modeling techniques to military aviation mishap reduction. There are numerous opportu-

nities for future research that can expand upon or develop from this research. We present a few of those potential opportunities here.

5.2.1 Application of AMIP to Navy and Army Aviation Mishap Data

USAF aviation mishap data from AFSAS is the most complete and best standardized of the aviation mishap data available across the DOD. The research team chose to perform this initial application of predictive modeling with the best data available. Future research could modify or expand AMIP as required to afford its predictive capabilities to the U.S. Army Combat Readiness Center (USACRC) and U.S. Navy Safety Center (NAVSAFECEN) military aviation data sets. NAVSAFECEN recently transitioned to a data system similar to AFSAS, which may result in easy application of AMIP to U.S. Navy (USN) aviation mishap data.

5.2.2 Expansion of AMIP Capability

AMIP in its current form is built to only fit and forecast Class A and B mishaps given individual mishap and TFH data. It is not capable of “drilling down” to aircraft type or model, or mishap cause. Such “slicing” of data for deeper analysis must be performed by USAF safety personnel or analysts by filtering report results in AFSAS, then loading that filtered data into AMIP. This process is explained in deeper detail in Section 3.7. Future research could expand the capabilities of AMIP to allow slicing of data loaded into it by MDS type or model. Future research could also build in keyword search capability so that models could be built using matches to those keywords that exist in the mishap One-Liner summaries. Source code for AMIP is available by request from Professor Ruriko Yoshida.

5.2.3 Application of TSA Modeling Methodologies to Other DOD Safety Programs

The incidence and causes of ground mishaps are not as well understood as aviation mishaps. Future research could attempt to apply TSA methodologies to military ground mishaps. The application of this research to ground mishap data could have a significant impact on the U.S. Army (USA) and U.S. Marine Corps (USMC), whose land combat centric mission training naturally results in the occurrence of ground mishaps, and could save the lives of Soldiers and Marines.

5.3 Conclusion

The research team hopes that these conclusions assist the USAF and DOD in developing more proactive safety programs. We believe this research can easily integrate into existing safety program processes across the DOD and will assist Commanders and safety professionals in targeting military aviation mishaps in a way that will save lives.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

- Federal Aviation Administration (2021) Operations network (OPSNET) total operations data. Accessed June 1, 2021, <https://aspm.faa.gov/opsnet/sys/main.asp>.
- Hyndman R (2006) Another look at forecast-accuracy metrics for intermittent demand. *FORESIGHT: The International Journal of Applied Forecasting* 4(4):43–46, <https://robjhyndman.com/papers/foresight.pdf>.
- Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts, Melbourne, Australia), <https://otexts.com/fpp2/>.
- International Civil Aviation Organization (2022) ICAO methodology for accident rate calculation and trending. SKYbrary. Accessed March 24, 2022, <https://skybrary.aero/articles/icao-methodology-accident-rate-calculation-and-trending>.
- Jaitley U (2018) Why data normalization is necessary for machine learning models. Medium. Accessed April 2, 2022, <https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029>.
- Kreisher O (2006) Toward zero mishaps. *Air Force Magazine* Accessed March 20, 2022, <https://www.airforcemag.com/article/1206mishaps/>.
- Li J, hang Cheng J, yuan Shi J, Huang F (2012) Brief introduction of back propagation (BP) neural network algorithm and its improvement. *Advances in CSIE* 2(AISC 169):553–554, https://link.springer.com/chapter/10.1007/978-3-642-30223-7_87.
- Light T, Hamilton T, Pfeifer S (2019) Trends in U.S. Air Force aircraft mishap rates (1950–2018). Technical report.
- National Commission on Military Aviation Safety (2020) Report to the president and the Congress of the United States December 1, 2020. https://www.militaryaviationsafety.gov/report/NCMAS_Final_Report.pdf.
- Office USGP (2019) Department of Defense aviation safety mishap review and oversight process. *Hearing Before the Subcommittee on Tactical Air and Land Forces of the Committee on Armed Forces, House of Representatives, 115th Congress, Second Session* (Washington, DC), <https://www.govinfo.gov/content/pkg/CHRG-115hhrg33381/pdf/CHRG-115hhrg33381.pdf>.
- Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (2018) Mishap notification, investigation, reporting, and record

- keeping. Technical Report DOD Instruction 6055.07, Washington, DC, <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/605507p.pdf>.
- Samaha K (2021) Aviation accident database & synopses. Accessed March 30, 2021, <https://www.kaggle.com/khsamaha/aviation-accident-database-synopses>, <https://doi.org/10.34740/KAGGLE/DSV/3021787>.
- United States Air Force (2020) Designating and naming defense military aerospace vehicles. Technical Report Air Force Instruction 16-401, Washington, DC, https://static.e-publishing.af.mil/production/1/af_a8/publication/afi16-401/afi16-401.pdf.
- United States Air Force Safety Center (2019) DOD human factors analysis and classification system (hfacs) 7.0 guide. Technical report, [https://www.safety.af.mil/Portals/71/documents/Human%20Factors/DoD%20HFACS%207.0%20\(AFSAS\)06-06-110309-983](https://www.safety.af.mil/Portals/71/documents/Human%20Factors/DoD%20HFACS%207.0%20(AFSAS)06-06-110309-983).
- United States Air Force Safety Center (2020) The U.S. Air Force mishap prevention program. Technical Report Air Force Instruction 91-202, Washington, DC, https://static.e-publishing.af.mil/production/1/af_se/publication/afi91-202/afi91-202.pdf.
- United States Air Force Safety Center (2021a) Mishap, HFACS, total flying hour, and number of active pilots data provided to the author via email, 1 September, 2021.
- United States Air Force Safety Center (2021b) Safety investigations and reports. Technical Report Department of the Air Force Instruction 91-204, Washington, DC, https://static.e-publishing.af.mil/production/1/af_se/publication/dafi91-204/dafi91-204.pdf.
- Wijaya CY (2020) The 80/20 split intuition and an alternative split method. Toward Data Science. Accessed March 20, 2022, <https://towardsdatascience.com/finally-why-we-use-an-80-20-split-for-training-and-test-data-plus-an-alternative-method-oh-yes-edc77e96295d>.
- Zhu Z, Ren B, Zhang X, Zeng H, Xue T, Chen Q (2021) Neural network-based probability forecasting method of aviation safety. *IOP Conference Series: Materials Science and Engineering* 1043(3), <https://doi.org/10.1088/1757-899X/1043/3/032063>.

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California