Air Force Institute of Technology
AFIT Scholar

Theses and Dissertations   Student Graduate Works

3-26-2015

Non-Rated Air Force Line Officer Attrition Rates Using Survival Analysis

Jill A. Schofield

Follow this and additional works at: https://scholar.afit.edu/etd

Part of the Operational Research Commons

Non-Rated Air Force Line Officer Attrition
Rates Using Survival Analysis

THESIS

MARCH 2015

Jill A. Schofield, Major, USAF

AFIT-ENS-MS-15-M-128

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-15-M-128

NON-RATED AIR FORCE LINE OFFICER ATTRITION RATES

USING SURVIVAL ANALYSIS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Jill A. Schofield, B.A., M. A.

Major, USAF

MARCH 2015

NON-RATED AIR FORCE LINE OFFICER ATTRITION RATES

USING SURVIVAL ANALYSIS

THESIS

Jill A. Schofield, B.A., M. A.
Major, USAF

Committee Membership:

Dr. Raymond R. Hill
Chair

Lt Col Matthew J. Robbins, PhD
Member

AFIT-ENS-MS-15-M-128

Abstract

The Air Force structures its workforce around rank structure and work specialty codes (Air Force Specialty Codes (AFSCs)). The challenge is to develop and manage personnel to fill a variety of skill sets at a variety of ranks over a 20-30 year planning horizon. To ensure that the missions are accomplished while adhering to congressionally-mandated force allocations, the Air Force is continually attempting to "right size" its force by maintaining the correct balance of personnel in each career field. The Air Force conducts its force structure management responsibility by comparing historical attrition rates to current manpower requirements for each AFSC to determine the "optimal" number of officers needed in each accession yeargroup over a 30-year career. Personnel analysts aggregate the individual yeargroup numbers for each AFSC and call this a "sustainment line." In this study, logistic regression was used to determine which factors are significant to predicting non-rated Air Force line officer retention. The variables considered were commissioning yeargroup, gender, source of commission, number of years served as enlisted, career field grouping, and distinguished graduate at commissioning source; all six were significant. All of these factors are included in the survival analysis, which yielded a total of 99 unique survival functions to characterize officer attrition behavior. Each of the survival functions provides a more specific representation of historic behavior that can be used to predict and/or shape future behavior. To best present the data to decision-makers, the unique survival functions must be aggregated after being weighted according to the respective percentage of the populations they represent.

This research is dedicated to my two amazing sons.

Acknowledgements

I would like to thank my advisor, Dr. Raymond R. Hill for his guidance, understanding, and most of all, patience. Although I often am, you never made me feel like the dumbest person in the room. I would also like to express gratitude to Lt Col JD Robbins, PhD, who provided insight and guidance throughout the journey–not only about this thesis, but also officership. Finally, Maj Jennifer Geffre, PhD, a special thank you for keeping me as sane as I can be while at AFIT.

Jill A. Schofield

Table of Contents

List of Figures

List of Tables

NON-RATED AIR FORCE LINE OFFICER ATTRITION RATES

USING SURVIVAL ANALYSIS

I. Introduction

1.1 Problem Background

Unlike commercial corporations, the size and rank composition of the US Air Force, as well as the rest of the Department of Defense, is dictated by law. Further complicating human resourcing, all Air Force members–enlisted and officer–start at entry-level. Personnel entering the Air Force inventory are said to be part of a particular yeargroup, labeled by their calendar year of entry. The challenge is to grow and manage each yeargroup to fill a variety of skill sets at a variety of ranks over a 20-30 year planning horizon.

Relating most population demographic-type studies to military personnel applications is particularly difficult due to the constraints placed upon the military by various laws, as well as the "military culture." Congress dictates to the services their maximum allowed population each year, whereas corporations are not bound by the same constraints. Corporations can grow or shrink themselves according to their workload and self-imposed profit margins. Generally speaking, the military cannot compensate for shortfalls in senior leadership by hiring someone from another corporation since the requirements of the job are built on the many years of experience it took to achieve that position. Corporations, on the other hand, are notorious for hiring executives from other companies to fill vacancies. Personnel decisions that the Air Force makes today affect manpower levels for up to thirty years, so these decisions

are not to be taken lightly.

The Air Force, like all the other services, structures its workforce around rank structure and work specialty codes (Air Force Specialty Codes (AFSCs) for the Air Force). These AFSCs define the skill sets required in the organization such that the organization can achieve its assigned mission. A "career field" is typically comprised of an officer AFSC and any related enlisted AFSCs and is managed by the Air Staff at Headquarters Air Force. Management includes career progression, as well as recommendations about force sizing.

To ensure that the missions are accomplished while adhering to congressionally-mandated force allocations, the United States Air Force is continually attempting to "right size" its force by maintaining the correct balance of personnel in each specialty code. Officers and enlisted receive valuable (and expensive) training to best fit into their career field and perform to the level required and expected. Additionally, the experience they gather throughout their careers is invaluable to the success of the Air Force and cannot be taught in a classroom. This type of investment must be carefully orchestrated to ensure that the resource outlay is not squandered simply because there happens to be an excess of personnel. This current research focuses on the officer corps and, unlike any civilian institution, the military cannot recruit mid- and senior-level leaders from outside the organization–they must be grown organically.

The Air Force conducts its force structure management responsibility by comparing historical attrition rates to current manpower requirements for each specialty (AFSC) to determine the "optimal" number of officers needed in each accession year-group over a 30-year career. Personnel analysts aggregate the individual yeargroup numbers for each AFSC and call this a "sustainment line;" this line indicates to personnel management decision makers how many Airmen are needed from each yeargroup to sustain the career field over a 30-year period.

Sustainment lines are used to make many personnel decisions. First, these lines are used to determine how many officers of each career type the Air Force should commission. By accessing the right number of lieutenants into each career field, the stage is set to fill the current definition of what each career field's requirements will be over the next 30 years. Additionally, sustainment lines are used to "right size" the number of officers when congressional mandates change. Since 2005, the Air Force has been scaling down its inventory of officers to meet reduced authorization levels due to the operations in the Middle East drawing to a close. To determine which career fields and yeargroups to trim, personnel management decision makers reference the sustainment lines, which were calculated using the target end-strength numbers. These sustainment lines are also used to identify which career fields and yeargroups may need retention incentives to ensure inventory does not drop to a level which would endanger mission accomplishment. Theoretically, the sustainment lines could also be used to invoke a "stop loss" policy to prevent eligible individuals from leaving the service.

Sustainment lines can affect national security and absolutely do affect every officer in the Air Force. Thus, it is imperative that they are as close as possible to reality. Unfortunately, reality changes every day. In the Air Force, new weapon systems are introduced and old ones are retired. Policies change how processes are performed, which changes personnel requirements. Organizations are realigned, combined, and separated, which causes additional adjustments to manning levels. Career fields are added, discontinued, combined and separated. Not even the Secretary of the Air Force can determine what the precise personnel requirements will be in the next ten years, let alone 30 years; however, projections must be formed, personnel manning decisions must be made, and both need done using the best information available despite our knowledge the projections are likely wrong (see [1] for a discussion of

Department of Defense prediction).

1.2 Research Scope

This thesis examined the process by which sustainment lines are determined and used. The research goal was to offer a reproducible, usable, accurate, and easily updatable model to be used in the sustainment line process. The work involved four phases. The first phase reviewed and reported on any past studies conducted in the field of military personnel management. The second phase defined the current process by which the sustainment lines are determined. Phase three statistically investigated the factors that affect an officer's career length and the last phase used those factors to develop a model that characterizes officer attrition and retention.

The majority of personnel studies conducted on Air Force officer retention tend to focus on rated officer (pilot, navigator, air battle manager, etc.) retention due to the large amount of money and time invested in their training. This study considered the non-rated line officers. The service commitments and career paths tend to be relatively equal within this group, so the attrition behavior was expected to be approximately the same. Since the full officer corps was considered a population, the group of focus in this study is considered a subpopulation.

1.3 Issues, Needs and Limitations

Headquarters Air Force Directorate of Personnel (HAF/A1) provided extracts from the Military Personnel Delivery System (MilPDS), the personnel database containing all active duty personnel records, covering 1999 to 2013. Over this time period provided, multiple career fields combined while others split; therefore, the analysts at HAF/A1 developed code to "modernize" older data and fill in any missing data points. The models created in this research are based on the modified data, and

therefore inherit the same assumptions that HAF/A1 made when "correcting" the data. These assumptions are discussed in Chapter 3.

A self-imposed research limitation was to implement all algorithms using SAS and Excel since the goal of this study was a reproducible, usable, accurate and easily updatable model to be used by HAF/A1 and its supporting agencies. The model was developed using the software that is most prevalent in those work areas–SAS and Excel. Most personnel analysts in the Air Force use SAS to mine and manipulate data, and export the results into Excel for final formatting or graphing. SAS is a powerful statistical software package that is capable of automating extremely complex algorithms and performing calculations on large data sets.

1.4 Thesis Outline

Chapter 2 reviews the literature surrounding military personnel inventory management, revealing various methodologies used and the different factors affecting career length characterization. Chapter 3 explains the data source, MilPDS, and the extracts provided by HAF/A1. Chapter 4 describes how the current sustainment model is calculated, as well as its assumptions. Chapter 5 discusses the logistic regression analysis (logit) and its findings. Chapter 6 details how the findings from logit were applied to create an attrition model. Chapter 7 details the assumptions and limitations of the attrition model, as well as recommendations for future research.

II. Literature Review

2.1 Introduction

Over the past fifteen years, multiple studies have explored the retention trends of military officers, as well as personnel management in different types of organizations. We review this body of research from two different angles: methodologies used and factors affecting retention.

2.2 Modeling Techniques

The body of research encompassing military (or military-like) personnel inventory management involves a wide range of methodologies from simulation to linear and logistic regression to stocks and flows (often used in system dynamics). "A simulation is the imitation of the operation of a real-world process or system over time. Whether done by hand or on a computer, simulation involves the generation of an artificial history of a system and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system" [2]. Software packages like Arena [3] and Simio [4] assist analysts in creating complex models of real-world situations, while less complex models can also be created in Excel, other programming languages, or remain conceptual. Analysts can track and collect different statistics during and after a simulation to quantify the effects of changes induced to the system simulated. As with any model, analysts make assumptions regarding how the system operates or how the entities (objects moving through the system) behave. "Once developed and validated, a model can be used to investigate a wide variety of 'what if' questions about the real-world system" [2].

One popular simulation paradigm is discrete-event simulation. "A discrete system is one in which the state variable(s) change only at a discrete set of points in time. The

bank is an example of a discrete system: the state variable, the number of customers in the bank, changes only when a customer arrives or when the service provided a customer is completed" [2]. For personnel management models, the discrete event might be entry or exit decisions or promotion events among the personnel entities. Discrete-event simulation results are analyzed to evaluate the system's performance. Changes can be made to the system in order to determine the best configuration for the stated objectives.

A specific type of simulation draws inspiration from complexity theory and is called agent-based modeling. "Agent-based modeling is a method for simulating the actions and interactions of autonomous individuals (the agents) in a network, with a view to assessing their effects on the system as a whole. Agents... actively make decisions, retain memory of past situations and decisions, and exhibit learning" [2]. The agents in this type of model, a special case of which is a complex adaptive system, abide by a set of rules that dictate how they respond to each other and their environment. The purpose of this type of simulation is to study the pattern of behavior the agents display as the simulation continues. A particularly useful behavior pattern is emergent behavior [5].

Regression analysis is also used to study military personnel inventories. "Regression analysis is a statistical technique for investigating and modeling the relationship between variables" [6]. The response variable from n factors is plotted in n dimensions and the linear relationship between the variables is determined by assigning a "best fit" equation to minimize the estimation error. The accuracy of the regression equation is determined by quantifying the prediction error of the model or the proportion of data variability explained by the model. In two-dimension linear regression, a straight line is drawn through the data and the response is typically a real number (as are the equation's coefficients).

Logistic regression analysis, or logit, is a regression model in which the response variable is binary instead of discrete or continuous. The model coefficients are typically continuous values. For example, if one is studying a medical process, the binary response could be alive/dead. In military personnel inventory studies, the binary response is typically still in/got out. The logit yields a probability of being in either state.

Survival analysis is a set of statistical techniques used to analyze "positive-valued random variables. Typically the value of the random variable is the time to failure of a physical component. . . or the time to the death of a biological unit" [7]. Where survival analysis differs from other stochastic techniques is "censoring," or only deriving a fraction of information from each observation. According to Miller [7], the three types of censorship are: stopping the trial at the end of a predetermined period of time, stopping the trial after a predetermined number (or percentage) of failures, and random censorship. Miller [7] illustrates random censorship by using a medical trial as an example: "In a clinical trial, patients may enter the study at different times; then each is treated with one of several possible therapies. We want to observe their life-times, but censoring occurs in one of the following forms:" loss to follow-up (the patient moves out of the area), drop out (the patient refuses further treatment or side effects prevent further treatment), or termination of the study [7]. Military personnel studies could consider the number of years of commissioned service as a "lifetime" and apply survival analysis.

System dynamics are simulations using the concept of "stocks and flows" to model systems from a resource-based view. "Stocks and flows–the accumulation and dispersion of resources–are central to the dynamics of complex systems. A population is increased by births and decreased by deaths. A firm's inventory is increased by production and decreased by shipments, spoilage, and shrinkage" [8]. The model

is created by first determining the stocks involved in the situation. In the case of military personnel management, the stocks could be commissioned years of service or individual ranks. The flows could be the personnel who are promoted into each grade (or year of service) and those who choose to leave the military for whatever reason. An enhanced feature of system dynamics simulation models is the feedback loop, which can affect the flows through the system. "Though there are only two types of feedback loop [positive and negative], complex systems can easily contain thousands of loops of both types, coupled to one another with multiple time delays, nonlinearities, and accumulations. The dynamics of all systems arise from the interactions of these networks of feedbacks" [8]. Feedback for military personnel could be incentives to remain in the system (bonuses), incentives to leave the system (early separation pay, early retirement), or a bottleneck (more individuals qualified for a rank than there are authorizations). In the military, promotions could also be considered a feedback mechanism where those who are not promoted, generally speaking, have an incentive to leave the system [8].

Chi-squared automatic interaction detection (CHAID) is an algorithm "to predict the response behaviour [sic] of individuals as accurately as possible" by dividing "a data set in exclusive and exhaustive segments that differ with respect to the response variable. The segments are defined by a tree structure of a number of independent variables, the predictors. To each segment of individuals, CHAID assigns a probability of response" [9]. The branch with the highest probability (or lowest, depending on the purpose of the study) reveals the characteristics of the ideal subject. For military personnel studies, the branches with the lowest probability of remaining in the service may require incentives to increase retention and the branches with the highest probability may require incentives to increase attrition. The branches could be defined by gender, career field, religion, source of commission, and rank since they

9

are mutually exclusive and collectively exhaustive categories.

2.3   Methodologies

According to Hill, Miller and McIntyre [10], discrete event simulation is used by the military in nearly every facet of operational analysis–from developmental testing in labs, to training, to strategic planning and decision-making. "Some of the critical issues facing the military in the aggregate include: How to structure the military given the uncertainty of the future..." [10]. This thesis addresses this very task with respect to personnel levels given uncertainty of the future.

Simulation modeling, and in particular agent-based modeling, has been a favored methodology of late. In an agent model, each individual (or group of individuals) is modeled as an independent entity interacting with other agents and the simulated environment. In the model, the agents are often assigned a utility curve to model their affinity (or lack thereof) towards particular stimuli. As any agent responds to the stimulus, the other agents are affected by this choice and may "choose" to follow suit. Each of the agents maintains an individual identity, yet is influenced by surrounding agents and their environment.

Example stimuli are money, time, or some other factor that would alter that agent's behavior. Using the focus of this thesis as an example, the stimulus could be the civilian economy. As the national economy becomes more robust, the actors–in this case, Air Force officers–may tend to gravitate more towards civilian employment instead of remaining in the military. Once an officer makes this decision, they may potentially influence those other officers with whom they interact. That influence may cause similar behavior (leave the Air Force or "attrit") or opposite behavior (stay in or "retain"). This interaction dynamic is captured by agents in the model.

Complex adaptive systems and agent-based modeling were created for the purpose

of studying how model entities (pilots) interact with one another [2]. Gaupp created the "Pilot Inventory Complex Adaptive System (PICAS): An Artificial Life Approach to Managing Pilot Retention" as his Master's Thesis [11]. He used money and time-off as stimuli and allowed the simulation-user to determine the specific utility curves (the pilots' attitude towards each of the stimuli), as well as the amount of the stimulus to apply in the model. Gaupp presented a sensitivity analysis regarding pilot retention in the Air Force as it is affected by how large a pay gap the agents experience by remaining in the Air Force vice flying for a commercial airline, as well as the amount of free time each pilot is allowed. While insightful, the result was a characterization of outside influences on pilot retention behavior and was not meant to predict attrition rates.

Schneider and Somers [12] discussed General Systems Theory and complex adaptive systems as they relate to leadership. They first described General Systems Theory as a foundation of most of the work in the field of leadership studies then they "develop the implications of organizations as CAS of the definition of leadership and the leadership process" [12]. CAS are quite useful in studies of leadership since each entity is affected by surrounding entities. In an organization, leadership (by definition) guides and attracts the personnel to influential action, which will hopefully further the shared mission [12]. Complexity theory and CAS are not meant to provide concrete numbers on which to base specific personnel manning levels, so this methodology is not explored in this thesis.

A study of the Canadian Forces (military) performed by the Department of National Defence in 2010 used the Arena software package and created ACME, the Arena Career Modeling Environment. ACME adequately models the career progression of the Canadian Forces' enlisted corps from the end of initial training ("boot camp") through the most senior rank, Chief Warrant Officer [13]. Each of the entities trav-

eling through the simulation carry a set of attributes that affect its behavior. This model replaces the Generic Modeling Environment (GeM) and adds the capability to determine the number of seats required in the mandatory promotion-based training courses each member attends throughout his career. This particular analysis used entity-based Monte Carlo simulation, although "entity-based Monte Carlo methods can also have some disadvantages when interested only in aggregate effects, or where the specific peculiarities of distinct individuals are not pertinent" [13].

Hall [14] used survival analysis to develop a tool focusing on forecasting enlisted Marine Corps members' retention. Survival analysis is where "a subject is observed, in an origin state, for a duration or episode until that subject leaves the origin state through an event or is censored and cannot be further observed. The duration of the origin state or episode and those causal factors that may have caused the event are analyzed" [14]. He calculated different "hazard rates per occupational field with gender, race, [and] citizenship" [14] and concluded that each subgroup of Marines behaved differently over time and should be modeled using different hazard functions. This approach appears sound and is investigated in this thesis.

Relating non-military personnel studies to military personnel applications are difficult due to the constraints placed upon the military by various laws and the "military culture" that civilian corporations are not bound by. Senior military leaders must be promoted from within the system, not recruited from outside the organization and certain "time in grade" and specialty requirements limit the pool from which to promote. In addition to the many constraints placed upon military personnel management, decision makers must remain mindful of the changing environment and attempt to maintain some level of flexibility to accommodate potential changes to personnel levels (e.g., surges or draw-downs).

The next study addressed issues associated with personnel management in a con-

strained environment with special consideration given to experience gained in that field of study that cannot be "recruited" from other organizations. Collofello et al. [15] studied personnel manning in a software development firm from the perspective of a project manager. In the studied organization, multi-month, multi-tier projects are doled out to a team of computer programmers who work together to complete the task. Throughout the development process, some of the programmers leave the project and the project manager must decide whether or not to replace the leaving individual. Like any organization, there are individuals with varying levels of experience, which may affect whether or not they are replaced. Unlike the military, the company has the option of hiring someone (or moving them from another project) with more than the basic level of experience. The software process model produced by this group of analysts using system dynamics (stocks and flows) provides the decision maker an opportunity to see the effects of different decisions in order to make the one that is most beneficial for the company. The results of this analysis show that the "right" decision depends on the priority for the project: schedule or cost [15].

A study conducted by the National Imagery and Mapping Agency (NIMA) used the concept of stocks and flows in an effort to evaluate personnel requirements over years to ensure that adequate manning was available to meet mission needs [16]. A stock is a quantity–such as food in a grocery store–and flows are rates at which the inventory flows into or out of stocks. NIMA is also a government entity, and is subject to many of the same constraints as the Air Force–primarily with respect to budget. The specific mission of the agency also limits how much it can recruit individuals from outside the organization; the work is quite specialized in that it takes many years to gain the experience required to operate at higher levels in the organization. Additionally, in order to work at a higher-level government civilian pay grade, one must hold a position of at least as much responsibility for a specified amount of time

13

before being promoted into the higher pay grade, which is quite similar to military ranks. Parker and Marriott [16] felt that the process was best modeled through stocks and flows. Cost was the primary driving force behind the model since it is the main constraint to the pay-grade balance that NIMA's senior leadership must maintain within the organization. In addition, the methodology produced in this particular study, allows workforce management decision makers to "play" with the model in an effort to explore how different decisions would affect attrition and promotion rates within the organization. Of note is the inclusion of a feedback loop that "... may have a great impact on the maturation of senior personnel at the tail end of the resource spectrum" [16]. This methodology may be useful when studying Air Force officers.

The four services in the Department of Defense differ in many respects. The United States Army is probably the most near-peer organization to the Air Force with respect to officer rank progression. Both services are constrained in manpower by Congress and face the same challenges with respect to "growing" experienced senior leadership. Additionally, both services struggle to understand retention trends and how they are affected by decisions and policies enacted by the highest echelons of leadership within the services. Dabkowski, et al. [17] used discrete event simulation, to understand the effect of attrition on the Army. This study focuses on West Point graduates as they progress through their military careers as an indicator of how the most "talented" officers retain in the Army. Dabkowski, et al. employ three different model scenarios in order to examine how the officer pool's "quality" would improve under different conditions. Each model only considers a single "dimension of talent and a single career path" when modeling officers through their careers [17]. While the simulation and models are fairly simple, insights were provided to Army leadership which may impact some of their policies regarding promotion and attrition

[17]. Although retention levels (quantities) were considered in the study, the overall concern was with the quality of the officers retained. It was insightful research, but not necessarily comparable to the current research focus.

A majority of the articles investigating the various factors that affect attrition used logistic (or logit) regression techniques [18, 19, 20]. Logistic regression is linear regression with a binary (two-level) response. With respect to assessing various factors, the binary dependent variable is inevitably did/did not retain. The findings of those specific studies are discussed individually in the next section.

2.4 Factors

The majority of research conducted on military personnel manning levels focus on developing methods to increase retention rates–whether targeted to a specific subpopulation (e.g., career field) or generalized to the entire force. Although that purpose is not within the scope of the current research, many of these studies use factors that warrant consideration in this research due to their observed effect on attrition. Some of the authors relied upon subjective data collected through questionnaires. Subjective data may induce a large amount of variance into the system and is difficult to reproduce. Since a goal of this thesis is a model that can be refreshed regularly with current data reflecting the total force, those studies based on subjective data are not discussed at length. Past research generally falls into two groups: testing whether a particular factor affects retention or developing a list of factors that affect the retention rates of a subpopulation within a particular military service.

Single Factor.

Demirel [18] postulated that the method by which military officers are commissioned affects retention rates when observed at two different points in their careers:

after the initial service commitment is concluded and at the end of ten years of service. His study concluded that "the retention rates of officers commissioned through the five major sources differ substantially. However, the effect of commissioning source on the retention of officers at the end of minimum service requirements is not large" [18]. Demirel's study is unique because he evaluated all four of the military services–first as a whole, then as individual departments–using Logit regression modeling and found that although each military service has different results, all show a difference in retention based on commissioning source. Military Academy graduates incur a five-year initial commitment, while officers commissioned through Reserve Officer Training Corps (ROTC–a university-based program) and Officer Training School (OTS–a 12-week commissioning program) graduates are required to serve for at least four years. After this initial commitment, the difference in retention behavior between the officers from the three different commissioning sources may continue well into an officer's career, which would provide an indicator of attrition rates.

Perry [19] studied the effect of career field (Primary Military Occupational Specialty–PMOS) on Marine Corps officers' retention and promotion to Major (O-4) and Lieutenant Colonel (O-5). She used logistic regression and Cox Proportional Hazard models (a type of survival analysis) and highlighted the correlation between some PMOS fields and retention, as well as promotions: "the results indicate that PMOS has a statistically significant effect on whether an officer survives until 10 [years of commissioned service]" [19]. While the Marine Corps handles their officers' careers quite differently than does the Air Force, career fields could prove to be a significant factor in retention and warrant consideration in the current research.

Conzen [20] investigated whether a military-sponsored graduate education was significant in the retention of officers in the Navy. He looked at two different types of sponsorship: fully-funded (in-residence graduate programs at Naval Postgraduate

16

School) or partially-funded (using Tuition Assistance funds at a civilian institution outside of the normal duty day). He used logit regression to find out if the officers remained in the Navy past the educational service commitment incurred. He found that, "A funded graduate education does not appear to have a substantial effect on retention past obligated service lengths but it is true that the proportion of officers with funded Master's Degrees leaving the Navy is consistently lower than that of those who earn a Master's degree on their own or have only a Bachelor's Degree" [20].

Until 2014, the Air Force encouraged officers to obtain a Master's degree as soon as possible in their careers in an effort to make them "competitive" for promotion. This unwritten policy resulted in a disproportionate amount of officers with graduate degrees–to the point where most officers have a Master's degree by the time they are a Major. Since the Air Force sends officers to AFIT, this factor warrants consideration in the current research. (However, it is likely that the minute population of Air Force officers with AFIT degrees will not provide a pool of officers large enough to be investigated given the parameters in the provided data set.)

Multiple Factors.

In addition to investigating how a single factor affects retention, studies have been conducted that explore how multiple factors affect attrition rates. Hall [21] used sequential logistic regression to develop a discrete-time logit model for Army dentists. The sequential logistic regression adds one variable into the system at a time and builds a new model with each iteration. Age, gender, race, dependents, commissioning source, residency completion, accession before or after October 2001, and deployments were explored, and in the final model (the eighth one), all factors except gender and deployments were included as significant indicators of retention.

Of note, he found that "all other factors held constant, the odds of Army dentists in the sample with dependents staying in the military is 56% greater than those without dependents" [21]. The data provided for this research does not contain reliable deployment or dependent data and are not considered; however, age, gender and commissioning source are examined to determine if they affect Air Force officer retention.

Castro and Huffman [22] took a different approach when studying retention rates of 289 US Army soldiers stationed in Italy and Germany. They surveyed the soldiers asking questions about operations tempo, work climate, leadership, family issues, and career intentions while also tracking the service member's years of service, rank, gender, ethnicity and age. They used CHAID to interpret the survey data then logistic regression to build models which predicted whether or not each individual intended to remain in service. The overall findings were that retention models need to include behavioral factors, as well as demographic data [22]. The response variable for the regression analysis was subjective data that may or may not have been taken seriously by the survey respondents. While the demographic data that Castro and Huffman used as part of their model are relevant to this thesis, the survey data that they collected are not.

Gjurich [23] asserted that Naval Surface Warfare Officers were leaving the Navy at too high a rate to ensure mission completion at higher echelons of leadership. He attempted to validate a conceptual model which considered responses from a survey, as well as demographic data collected through the Navy personnel office. The model used logistic regression on data from Navy Lieutenants (O-3s) who had already finished their initial service commitment. He considered source of designation (regular or reserve commission–no longer a consideration in the military), commissioning, dependent status, level of education, yeargroup and race and found that all but

dependent status, yeargroup and race were significant [23].

Zinner [24] compared data from 1992 to that from 1996 regarding male Marine Corps officers in their initial commitment window to determine which factors affected their retention four years later once they were no longer bound to the service. He took a "broad social science approach combining organizational and individual behavioral factors" to build a logistic regression model of retention decisions [24]. He found that the demographics commissioning source, occupational specialty, and deployments as well as the service member's perception regarding job satisfaction, civilian job searches, perceived job security, perceived job transferability, and spouse's career were all significant [24]. Much of the data was collected using a survey then matched with data collected from the Defense Manpower Data Center.

III. Data Source

3.1   Introduction

Typically, the very first step in an analysis entails becoming familiar with the data. Prior author experience in this area precluded a lengthy study; however, a full explanation of the data is found in the next section. The purpose of this chapter is to provide background on the data for future studies in this area.

3.2   MilPDS

In the Air Force, all personnel data are stored in a database called the Military Personnel Delivery System (MilPDS). Each individual is allotted over 300 data record fields that can be populated throughout his/her career. These fields span a wide variety of data points: full name, identification number, home address, duty assignment (current, as well as a full history), AFSCs (current, as well as a full history), gender, rank, projected promotions, dates of service, military awards, flying hours, etc. The values for these fields are not updated by the individual, but by a trained personnelist or are automatically updated within the system (like "years of service" will automatically increment when an anniversary passes). The trained personnelists are located on each base and require documentation in order to change values that are already in the system.

A database as large as MilPDS with hundreds of personnel inputting data from all over the world 24 hours a day is an ever-changing and sometimes unstable program. Data back-ups are saved frequently to minimize the impact of a malfunction. The data (as a whole) is never up-to-date since changes are constantly input.

Occasionally, records are incomplete or incorrect. While a multitude of reasons exist, most records problems are due to human error. Any time data are manually

entered into the database, there exists the risk of mistakes. If an individual is new to the military, the personnelist creating the new record might leave a field blank or incorrectly input data. When updating a record, the personnelist may input the new data incorrectly or unintentionally alter pre-existing data. Some updates require the member to initiate the change to MilPDS. For instance, if someone earns an award on a deployment, he/she is required to take the documentation to a personnelist to update the record. If the member waits (or fails) to have the record updated, the database will be inaccurate. When database maintenance is performed, data can also be lost (although it can typically be recovered by using a back-up). Most of the incomplete or incorrect entries are eventually corrected.

3.3 Extracts Provided by HAF/A1

The data used in personnel analyses at HAF/A1, and its supporting agencies, are actually extracted from MilPDS, consistently created at the end of each month at around the same time of day. These extracts include the vast majority of the fields found in MilPDS, but not all of them. If the extract is generated between the time an error is input to MilPDS and when it was corrected, then the extract will only contain the errant data. These extracts are also saved in SAS format for ease of use and are often referred to as "snapshots" since the database is constantly changing.

As with any analyses, the results are only as accurate as the data used to calculate them. The analysts at HAF/A1, as well as its supporting agencies, are well-versed in many of the shortcomings of the extracted data and have developed programs to automatically fix some of the errors. For instance, if an individual's service date is missing, the program will automatically scan previous months' extracts for the missing value in case it was inadvertently deleted. Also, since many analyses are performed for a specific AFSC, it is important that the field they are referencing is

not blank and is as accurate as possible. If this field is missing, the program knows to look at other AFSC fields and duty history fields to fill in the "best guess".

Throughout an officer's career, he can accumulate multiple AFSCs. The Core AFSC (or Core Identifier, Core ID) is a 3-character code that reflects the career field to which the officer belongs and can only be changed if "the officer formally applies and is approved to retrain is designated for involuntary cross flow or is approved to transfer to another competitive category" [25]. The "primary AFSC" is defined by Air Force Instruction 36-2101 as "the AFSC...in which the individual is most qualified to perform duty" and is based on skill/qualification level, experience, complexity of the specialty, amount of formal education and training and currency of equipment qualification [25]. If an officer has additional qualifications, he can be awarded second and third AFSCs, "in order of best qualifcation" [25]. Whenever the officer reports to a duty assignment, his records are updated with the "duty AFSC" corresponding to the position he holds on the unit manpower document (UMD) [25].

With the exception of Core IDs, officer AFSCs can be up to 6 characters long. The first character is a "prefix" and reflects special duties "when there is a need to identify an ablity or skill not restricted to a single utilization field or career field" [25], such as commander, trainer, or flight-qualified (for non-rated career fields). The next three characters correspond to a Core ID or special duty identifier, such as executive officer, instructor, recruiter, or student. The fourth character is a numeric representation of the skill level of the position. While each career field defines the specific requirements for this digit, most use a "1" for entry-level, "3" for non-entry level, and "4" for staff positions at the Wing level or higher. The last character, if present, is often called a "shred" or suffix and refers to a specialty within a Core career field [25]. For instance, prior to 2010, the 61S ("Scientist") career field had four shreds: (A) Analytical, (B) Behavioral, (C) Chemist, and (D) Physicist.

Typically, in officer personnel analysis, statistics are calculated by career field, as determined by an officer's Core AFSC. Sometimes officers perform duties outside of their main career field (instructor, student, etc.) and their Duty AFSC will reflect that, but the Core AFSC does not change; therefore, this field is used to filter or categorize the data. Pilots and navigators (all of them officers) do not generally have a value in the Core AFSC field because they are identified by the type of airframe in which they fly, which is designated by a Rated Distribution and Training Management (RDTM) code. To make the coding for analysis involving rated officers easier, the personnel analysts have created programs that "convert" RDTM codes and fill in the Core AFSC field. Unfortunately, this is not a short code–akin to a "decoder ring"– since most aircrew have flown in multiple different airframes (especially during pilot training).

For this research effort, HAF/A1 provided monthly extracts of personnel data from January 1999 through December 2013 that contained data for Active Duty Air Force Officers. Within this period of time, the number of officers in the extracts fluctuated between 63,500 and 73,100 and each one had between 315-360 fields of data (many of which are blank). For this study, the data sets were filtered to include only non-rated, line officers, which means that rated officers (pilots, navigators, air battle managers, etc.), medical officers (doctors, nurses, medical logistics and administration, etc.), chaplains, and attorneys are not included. Both the factor determination and model development required different subsets of the extracts, which are described in their respective sections.

IV. Current Sustainment Model

4.1 Introduction

HAF/A1 provided the SAS code they currently use on personnel data to develop the sustainment lines for each career field in the Air Force. The code, originally written in 2007, is over 100 pages when printed and includes little documentation about the specific calculations being performed. Since 2007, the code has been updated and appended to reflect changes in the AFSC structure, as well as other Air Force policy changes. The code currently in use performs a maximin-flow optimization over all career fields to determine each career field's sustainment line [26].

4.2 Core AFSCs

Over half of the SAS code is geared towards ensuring that the AFSCs for each individual (especially rated officers) are correct [26]. Over time, career fields merge and split based on Senior Leaders' direction. To ensure that the most current AFSC structure is reflected in the results, historic data are updated to reflect various merges and splits as they apply. Additionally, the SAS code relies upon each officer's "core AFSC," which is the three-character specialty code that reflects the career field to which the officer belongs, regardless of which job they are performing at the time (or "duty AFSC"). For instance, if an engineer (core AFSC 62E) is serving in an instructor position, he may have a 92T duty AFSC ("Instructor," which is generic with regard to career field) or T62E, which indicates he is teaching within the engineering career field. This officer should be categorized as a 62E for the purpose of developing the sustainment code.

The rated career fields typically do not populate the core AFSC field for their officers because they are identified using a rated distribution and training management

(RDTM) code, also referred to as "aero ratings". Each airframe in the Air Force inventory is assigned a unique code and each hour rated officers operate each airframe is recorded in their records, which is how the rated career field managers track their officers. In order to apply the sustainment code to these officers, their core AFSC field must be populated. The SAS code provided by HAF/A1 is largely focused on ensuring this is done properly. The code takes into consideration RDTM codes, aviation service codes ("flying hours"), and duty AFSCs to populate the core AFSC field for rated officers.

4.3 General Methodology

The sustainment code used by HAF/A1 to derive sustainment lines typically uses empirical data to formulate retention and crossflow trends categorized by core AFSC and prior enlisted service [26]. The SAS code provided by HAF/A1 shows that annual data from 1999 through 2012 (the last completed fiscal year) were used, and data from three years (2006-2008) were excluded for retention rates due to involuntary force shaping efforts. The raw retention percentages are used "to build a large interconnected network of nodes that is assumed to emulate how varying types of officers will flow through the Air Force system ove the next 30 years. Using this network, a maximin-flow optimization is invoked to maximize the projected manning of the lowest manned career field in 30 years. This optimization is subject to several constraints, surch as the congressionally mandated end-strength level on the five-year development plan" [26].

The current model formulates sustainment behavior for each core AFSC, which determines the shape of the sustainment line but leaves the height (or y-intercept) undetermined. Given the funded manning reqirements (the positions that the Air Force has determined are essential to operations, typically equal to the congressional

mandate) for each career field, the sustainment lines are then optimized to allocate the "right" number of officers to each core AFSC by maximizing the number of officers in lowest-manned career field (determined by comparing the current inventory to the funded requirements).

To make projections for each career field out to 30 years, analysts need to make some assumptions. According to Gibb, there are three key assumptions in the current model. "The trends of historical retention, utilization, and crosslfow of officers will continue for the next 30 years; The five year out congressionally mandated end-strength will not change for 25 years; and The five year out funded [manning] requirements will not change for 25 years" [26]. Although these assumptions are wildly unrealistic, they are imperitive to creating the model. It is impossible to predict the future with any certainty and military manning requirements and allocations, as well as retention behaviors, are constantly changing. Congressional mandates with respect to military manning are shaped by the number and size of conflicts in which the military is supporting, as well as the President's fiscal priorities. The strength of the economy plays a significant role in whether officers stay in the military or leave to pursue more potentially lucrative careers in the civilian sector.

V. Analysis

5.1 Logistic Regression

Logistic regression was used to determine which factors are significant to predicting non-rated Air Force line officer retention. The predictor variables are categorical and the response variable is binary. The variables considered were commissioning yeargroup (the calendar year that the individual became an officer - "yeargroup"), gender, source of commission (Officer Training School (OTS), Reserve Officer Training Corps (ROTC), or any service Academy - "commission"), number of years served as enlisted (binned into 0-2, 3-4, 5-7, 8-11, or more than 11 years - "prior enl"), career field grouping as determined by the first digit of the Core AFSC (non-rated operations (NRO), logistics (LOG), base support (SPT), acquisitions (ACQ), and Office of Special Investigations (OSI) - "Career Field"), and distinguished graduate at commissioning source (typically the top 10% of graduates determined by academic grades, as well as military training - "DG").

Data.

The extracts from MilPDS were combined to create a cohort. In a cohort, each officer, identified by a unique serial identification number, was represented by one line of data that contained fields for gender, source of commission, etc. for each year of data used in this analysis (1999-2013). If the officer was not on active duty during any of the years covered by the data, those corresponding fields were empty. It was assumed that the most recent record was the most accurate, so "stagnant" variables (those variables that do not change over an officer's career) like gender, prior years of service, etc. were determined by looking at the most recent year the officer was in the Air Force.

The cohort was then divided into smaller data sets based upon the CCR(s) over which the officer's career spanned. For instance, if an officer joined the Air Force in 2000 (0 CYOS) and separated in 2012 (12 CYOS), then that officer's record would be included in the 0-6 and 4-8 CCR data sets, but not the 8-14 since the forteenth year is outside the span of time that the data covers. This is considered "truncated data." Including data not spanning the time frame would skew the data. Additionally, if an officer joined the Air Force in 1995 (0 CYOS) and was still in the Air Force in 2012 (17 CYOS), only his time captured by the provided data sets (1999 through 2013 or 4-17 CYOS) would be considered. This is called "left-truncated" data, and if not included would greatly reduce the amount of data available for calculation.

Each entry in the data sets included the officer's unique serial identification number, commissioned yeargroup, gender, source of commission, number of years of enlisted service (categorical), career field grouping, whether or not they were a distinguished graduate from their commissioning source, and whether or not they retained at the end of that CCR period. In each data set, there was exactly one entry for each officer whose career had the potential for spanning that CCR period given the provided data sets.

Calculation.

Each CCR data set was then processed through logistic regression using the 'proc logistic' function in SAS. The analysis of effects table provided a signficance estimate for each of the effects using a Wald Chi-Square statistic and a corresponding p-value. The Wald Chi-Square statistic "is the squared ratio of the Estimate to the Standard Error of the respective predictor. The Chi-Square value follows a central Chi-Square distribution with degrees of freedom given by DF, which is used to test against the alternative hypothesis that the Estimate is not equal to zero. The probability that a

particular Chi-Square test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is defined by" the p-value [27]. The smaller the p-value, the more significant the factor.

5.2 Findings

With only five exceptions (shaded), all six of the variables were found significant in each of the five CCRs at the individual 95% confidence level as shown in Table 1. All SAS code is included in Appendix A.

Table 1. Analysis of Effects Summary of P-Values.

CCR	Obs	Yeargroup	Gender	Commission	Prior Enl	Career Field	DG
$0-6$	20,789	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
$4-8$	22,834	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
$8-14$	11,487	< 0.0001	< 0.0001	0.3894	< 0.0001	< 0.0001	< 0.0001
$12-19$	8,154	0.9352	0.0083	0.0020	< 0.0001	0.1784	< 0.0001
$20-22$	7,386	< 0.0001	0.0835	0.0001	0.4479	< 0.0001	0.0002

Discussion.

Logistic regression indicated that yeargroup was significant in explaining retention in all CCRs except for 12-19 CYOS, illustrating that officer retention behavior changes over time–either voluntarily or involuntarily. Many force-shaping efforts target officers before the end of their initial commitment, so retention throughout the 0-6 CCR is expected to be lower for the yeargroups that faced force shaping. Reduction-in-force (RIF) boards target officers who have completed their intial commitment and who have not completed 16 CYOS. The difference in retention between yeargroups targeted by a RIF and those who have not are likely why 'yeargroup' is signficant in the 4-8 and 8-14 CCR data sets. Another force-shaping measure is the Selective Retirement

Board, aimed at officers with more than 20 CYOS, thus explaining the significance in the 20-22 CCR. The likely reasons that yeargroup is not significant in the 12-19 CCR is (1) when an officer finishes 12 CYOS, he has typically made a career decision and will probably choose to stay at least until completing 20 CYOS and (2) force-shaping efforts aimed at these officers, like the Selective Early Retirement Board (SERB), are used sparingly (often at the expense of other force-shaping measures). In addition to these involuntary seperation measures, officers may choose to leave the Air Force based on other factors like operations tempo, base closures, changes to their career field, or other policies that lead the officers to believe that the military is not taking care of them or that their career futures are dim.

Gender was a significant predictor for all CCRs except for 20-22 CYOS. In the other CCRs, males were 1.33 to 1.86 times more likely to remain in the Air Force than women, as illustrated in Figure 1. Odds ratios are calculated by SAS during the logistic regression procedure by exponentiating the parameter estimates from the logit regression model [27]. The odds ratio is interpreted as one setting for a variable being n times more likely to occur than the baseline, given that all other variables are held constant [27]. For example, in Figure 1, in the 4-8 CCR, males are a little more than 1.8 times more likely to retain in the Air Force than females, given that they are in the same yeargroup, career field, etc.

Although the actual reason cannot be determined based on this data, one can attribute the higher rate of female attrition to roles within the traditional family. Women likely commission when they are single and as time passes, they meet their mates, get married, and begin to have children. Some may feel compelled at that time to leave military service to shield their families from the instability induced by military deployments and moves.

Commissioning source was significant for all CCRs except for 8-14 CYOS. Figure
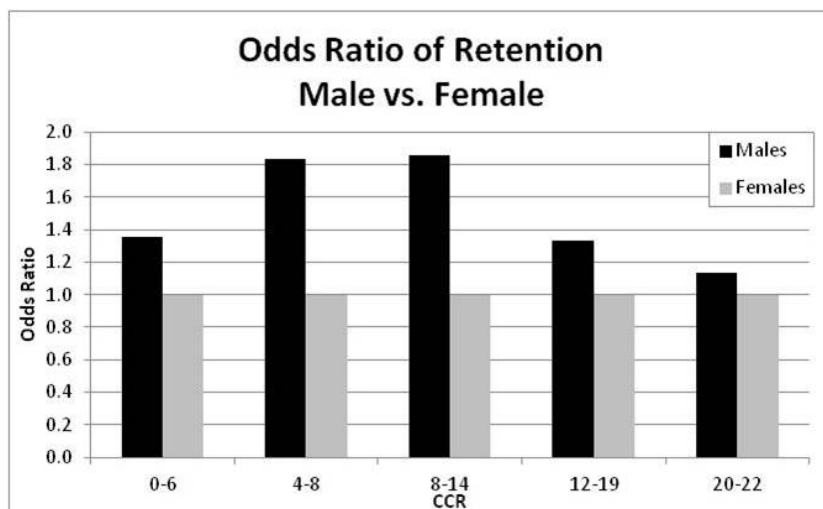
Figure 1. Odds Ratio of Retention for Gender

2 holds Academy as the baseline and compares the other commissioning programs to that rate. The odds ratios show that OTS graduates are almost 1.6 times more likely to finish 6 CYOS than the other two programs. This is likely due to the fact that a large proportion of individuals who earn their commission at OTS are prior enlisted, who are closer to retirement and thus more "invested" in their careers. The majority of the non-prior enlisted officers are typically individuals who had careers in the civilian sector and chose to join the military either because of an increased sense of duty to their country (especially those who joined as a reaction to the terrorist attacks on September 11, 2001) or because they were unhappy with their civilian job and were looking for a career in the military. OTS graduates are a little more than 2.21 times more likely to retain than Academy graduates (and 1.24 times more likely than ROTC graduates) at the 4-8 CCR. The Academy graduates' retention at the 4-8 CCR is significantly lower than the other commissioning sources. This is likely due to the fact that Academy graduates incur a 5-year service commitment, whereas OTS and ROTC graduates incur a 4-year service commitment. The 0-6 CCR captures OTS and ROTC graduates who separated after their initial commitment; however, the 4-8

31

CCR only captures Academy graduates who leave after their initial commitment (it also includes officers who left after their second or subsequent assignment, as well). The difference in service commitments explains the differences between the three commissioning sources at the 0-6 and 4-8 CYOS.
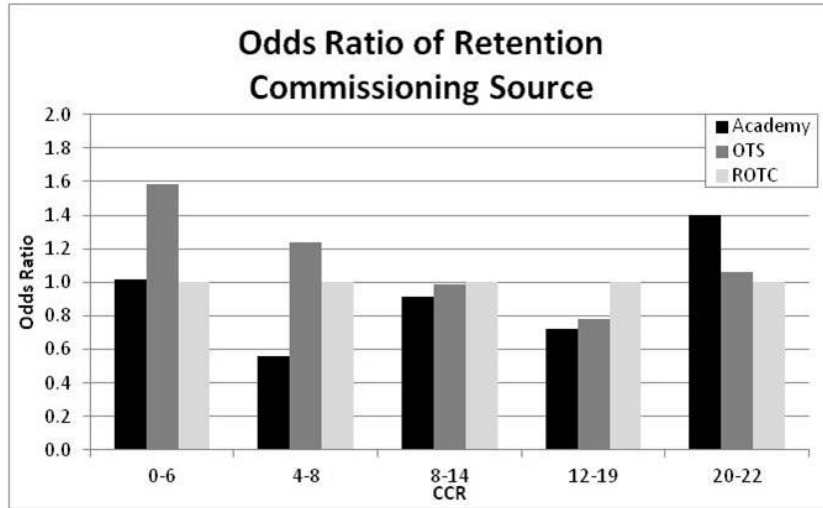
Figure 2. Odds Ratio of Retention for Commission Source

At 12-19 CYOS, OTS graduates (many of whom have prior enlisted time) attrit at a higher rate compared to ROTC graduates than in previous CCRs likely due to their eligibility for retirement. ROTC graduates have the highest attrition rate in this CCR. In the 20-22 CCR, Academy graduates retain at a higher rate than the other commissioning source, likely due to the perception that they have a higher probability of being promoted to the higher echelons of the military organization, such as into the General Officer corps.

The prior years of service predictor behaves as expected. Officers with more time in service tend to retain much better than those with fewer years until they reach retirement eligibility. Typically, military members with 20 years of service (enlisted and officer time combined) are eligible to retire. If an officer is prior enlisted, then he/she can retire once that milestone is reached; however, if they want to retire as an

officer, they have to accumulate at least 10 CYOS. With the recent force management efforts, the 10 year requirement was waived to 8 years and many eligible prior enlisted officers attrited under the program. Figure 3 uses "0-2 prior years of service" as the baseline. The 12-19 CCR provides evidence that prior enlisted officers generally retire at their respective 20 CYOS. Those officers who have more than 11 years of enlisted service are retirement eligible within the 8-14 CCR timeframe, which is likely why their retention is respectively low at that point.
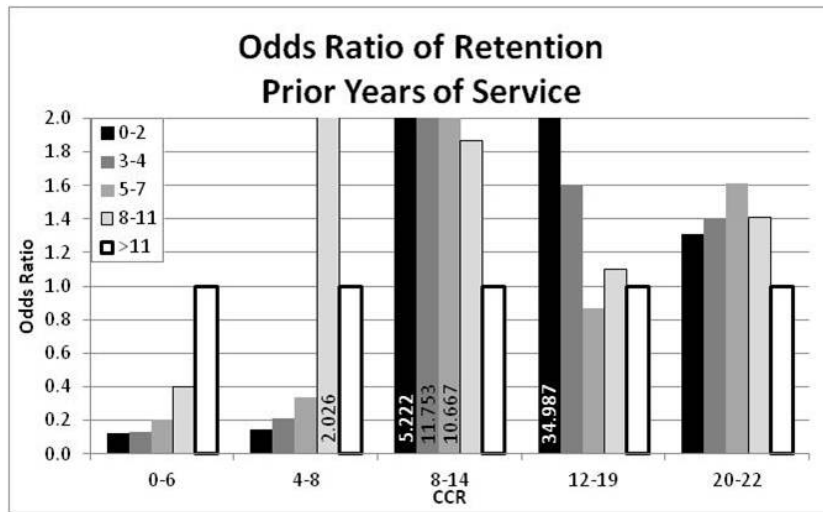
Figure 3. Odds Ratio of Retention for Prior Enlisted Years of Service

Career field groupings were significant to the logit regression model over all of the CCRs except for 12-19. Given the odds ratio calculations, the first conclusion was that OSI may be affecting the results (see Figure 4). As an excursion, OSI officers were removed from the data and logit regression was performed again. The results remained the same, leading to the conclusion that the other career fields behave differently from each other and OSI was not the only one that was different. Figure 4 illustrates the odds ratios of retention with support officers (3XX AFSCs) as the baseline showed no significant pattern in retention trends among the career fields.

Officers who were distinguished graduates (DG) from their commissioning source
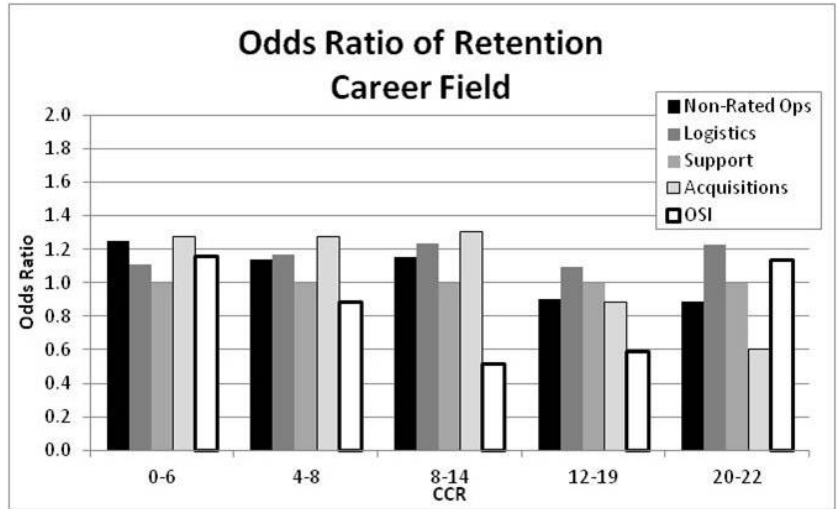
Figure 4. Odds Ratio of Retention for Career Field Groupings

retained 1.29-2.13 times better than non-DG officers over the entire span of this study, as illustrated by Figure 5. DG was the only factor that was signficant to the logistic regression for all CCRs. Since DGs are the top 10% of their graduating class, they are treated like the "best and brightest" the military has to offer since they attain the DG designation through direct competition with their peers.
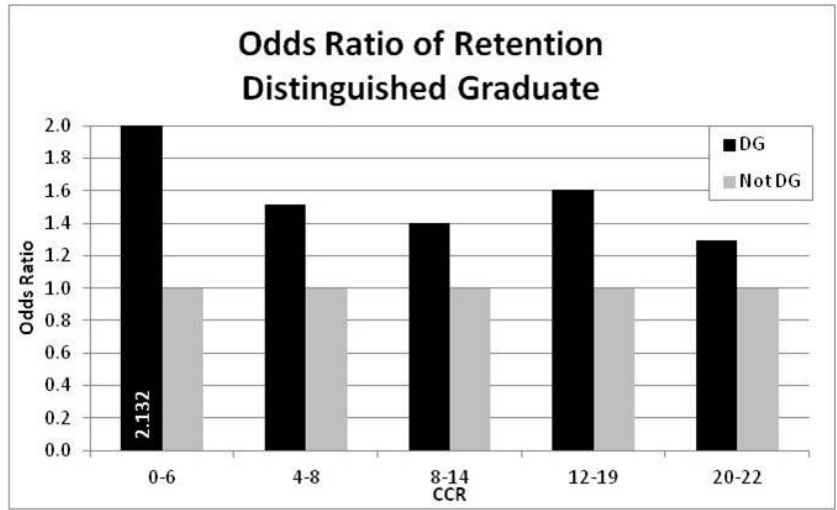
Figure 5. Odds Ratio of Retention for Distinguished Graduate

5.3 Summary

Logistic regression on each of the five CCRs found that all six of the factors of interest (commissioning yeargroup, gender, commissioning source, prior enlisted years, career field grouping, and distinguished graduate from the commissioning source) were significant. All of these factors are included in the sustainment model for non-rated line officers in the Air Force.

VI. Application

6.1 Survival Analysis

Survival analysis was used to estimate alternative estimates for non-rated line officers' retention behavior because "ordinary least squares regression methods fall short because the time to event is typically not normally distributed" [28]. Additionally, survival analysis accommodates "censored" data, which means that the data can no longer be collected, but the event of interest (in this case, attrition) was unable to be observed. Using survival analysis maximizes the amount of data available for model fitting analysis.

Proportional hazards regression is often used to analyze survival data and is easily performed with SAS software using the "proc phreg" function. "The Cox proportional hazards model assumes a parametric form for the effects of the explanatory variables, but it allows an unspecified form for the underlying survivor function" [29]. The explanatory variables used in this study are all categorical, which is why a parametric form is needed. An added feature in SAS 13.2 is the capability to allow for left-truncated data in addition to censoring. "Left-truncation occurs when individuals are not observed at the natural time origin of the phenomenon under study but come under observation at some known later time... Thus, any contribution to the likelihood must be conditional on the truncation limit having been exceeded" [29]. The provided data only covers 15 years, so following a single yeargroup's 30-year career is impossible; therefore, a methodology that considers data that empirically represents the behavior in the latter years of officers' careers is critical.

Figure 6 illustrates left-truncated and censored data as they relate to the timeline in which data were collected. The top bar shows an officer who was commissioned prior to 1999-the first year of data provided-and left the Air Force in 2003. This data

are considered left-truncated since the rest of the 1995 yeargroup's retention behavior is unknown until 1999. The second officer was commissioned in 2001 and left the Air Force in 2011, which means that the entire record can be considered. The third officer was commissioned in 2005 and was still in the Air Force during the last data set, so this information is censored since his overall attrition behavior cannot be determined with the provided data. The first and third officer in this illustration can still provide valuable retention inputs to the model when viewed with a conditional probability. The first officer's behavior, for example, can be interpreted as "given that this officer completed four years of service, he remained for four years." Similarly, the third officer's behavior can be interpreted as "given this officer commissioned, he stayed in the Air Force for at least eight years."
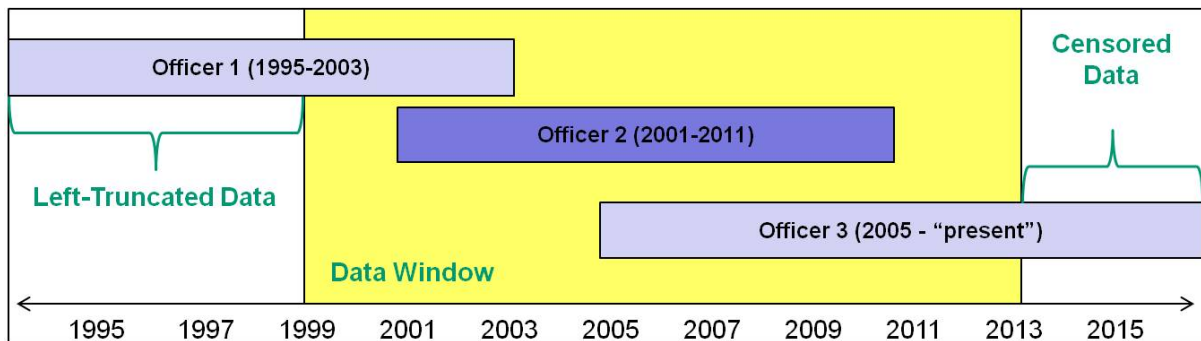
Figure 6. Illustration of Left-Truncated and Censored Data Using Notional Officers

Data.

The extracts from MilPDS were combined to create a cohort. In a cohort, each officer, identified by a unique serial identification number, was represented by one line of data that contained fields for gender, source of commission, etc. for each year of data used in this analysis (1999-2013). If the officer was not on active duty during any of the years covered by the data, those corresponding fields were empty. It was assumed that the most recent record was the most accurate, so "stagnant" variables

(those variables that do not change over an officer's career) like gender, prior years of service, etc. were determined by looking at the most recent year the officer was in the Air Force.

Start and stop variables were created and were populated with the officer's CYOS in the first and last dataset in which they appeared, respectively. These variables are used to characterize left-truncation and censorship when applicable. If an officer retrained into a new core AFSC ("crossflowed"), his AFSC retention was calculated as "no," but his Air Force retention was calculated as "yes." In this case, a new line of data was created so that the remainder of the officers' retention behavior is attributed to the new AFSC.

Each entry in the data sets included the officer's unique serial identification number, gender, source of commission, number of years of enlisted service (categorical), career field (determined by the first digit of the core AFSC), whether or not they were a distinguished graduate from their commissioning source, start CYOS, stop CYOS, and a censor variable.

Calculation.

The data were analyzed by career field using the Cox proportional hazards model by invoking the 'proc phreg' function in SAS (code included in Appendix B). The stepwise option was included to ensure that each career field's model only contained the factors significant to the regression and that the model was not fitting noise within the system. The technique was performed by career field and considered gender, source of commission, number of years enlisted, and distinguished graduate status and also considered that some of the data was left-truncated. In addition to the five career field groupings considered (non-rated operations, logistics, support, acquisitions, and Office of Special Investigations), the whole dataset (non-rated line

of the Air Force or "NRL") was also analyzed.

This calculation yielded a set of regression equations (one for each career field plus one for NRL) with varying numbers of factors and different coefficients, as illustrated in Table 2. Due to its small size, the OSI career field did not yield a significant regression equation and was hereafter not considered as a separate career field, but was still used in the NRL calculations.

Table 2. Coefficients Found Using Stepwise Proportional Hazards Regression.

Career Field	Gender (Female)	Academy Grad	OTS Grad	Enlisted 0-2 Yrs	Enlisted 3-4 Yrs	Enlisted 5-7 Yrs	Enlisted 8-11 Yrs	DG (No)
NRO	0.0853	-0.0144	0.0593	0.0697	0.2710	0.1509	0.0857	0.1573
LOG	0.1173	—	—	0.0970	0.2354	0.2544	0.0947	0.1577
SPT	0.1069	—	—	—	—	—	—	0.0923
ACQ	—	0.1026	0.1046	0.1457	0.2507	0.1990	0.2558	—
OSI	—	—	—	—	—	—	—	—
NRL	0.07612	0.0511	0.0504	0.1341	0.2651	0.2017	0.1372	0.0975

NRO - Non-Rated Operations (1XX)
LOG - Logistics (2XX)
SPT - Support (3XX)
ACQ - Acquisitions (6XX)
OSI - Office of Special Investigations (71S)
NRL - Non-Rated Line (1XX, 2XX, 3XX, 6XX, 71S)

These coefficients were then used as baseline covariates in another iteration of proportional hazards regression. Baseline covariates are "used to request a survival curve that represents the survival experience of an average patient in the population" which the coefficients represent [29]. This second iteration resulted in a set of survival functions for each career field, detailed for each commissioned year of service (CYOS) and includes only the variables found significant by the first iteration of stepwise proportional hazards regression. Each career field has a unique survival function for each setting of the significant factors found by the stepwise regression. SAS also calculates a 95% confidence interval for each function. All SAS code is included in

Appendix B.

6.2 Findings

To characterize non-rated line officers' retention probability by career field, 99 different survival functions are necessary (see Table 3) since each function is based on up to 60 different combinations of variable settings. For the sake of brevity, only one combination is discussed for each career field.

Table 3. Number of Survival Functions by Career Field.

Career Field	# Survival Functions
NRO	60
LOG	20
SPT	4
ACQ	15
NRL	60

Non-Rated Operations (NRO).

The non-rated operations (NRO) career field involves officers whose first digit in their core AFSC is a "1" and excludes pilots (11X), navigators (12X), astronauts (13A), air battle managers (13B), and attack remotely piloted aircraft (RPA) pilots (18A). NRO includes control and recovery (13D), air liaison officers (13L), airfield operations (13M), space and missiles (13S), intelligence (14N), weather (15W), and cyberspace operations (17D) officers. Stepwise regression revealed that prior enlisted service (5 categories), gender (2 categories), commissioning source (3 categories), and distinguished graduate (2 categories) were all significant variables in determining the survival function. Addressing each combination of these settings resulted in 60 different survival functions. Figure 7 shows the survival function for NRO officers with

no prior enlisted experience, who are male and graduated from any one of the Service Academies, but did not receive distinguished graduate honors from their commissioning source. This plot includes a 95% confidence interval for the function.
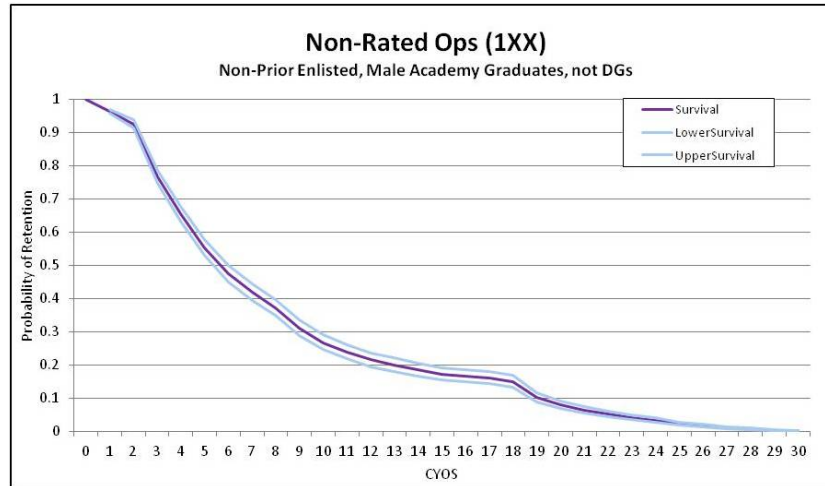
Figure 7.   Non-Rated Operations Survival Function (Non-Prior Enlisted, Male Academy Graduates, not DGs)

Each point on the line is an "instantaneous" retention rate, which is interpreted as the probability that the population currently investigated will retain to a particular point in time. Since all of the calculations were performed on a discrete number of years of service and the results are discretized on CYOS, the line connecting those points is an interpolation of the data. Air Force personnel analysts typically use the discrete data in their calculations and the line connecting those points primarily serve an asthetic purpose. The slope of the line connecting the points indicates increases or decreases in attrition rates over a period of time. Figure 7 reveals a steep decline in retention after 4 CYOS that levels out around the 10 CYOS point. Only 26.7% of the original NRO population illustrated is projected to still be in the Air Force at 10 CYOS and only 7.8% are expected to stay until 20 CYOS.

Of the 60 different survival functions for NRO officers, the best retention was found in male ROTC graduates with more than 11 years of enlisted service who earned a

DG from their commissioning source. At 10 CYOS, 34.3% are expected to still be in the Air Force and 12.8% are expected to remain at 20 CYOS. The worst retention of NRO officers was for females who went to Officer Training School (OTS) with 3-4 years of enlisted service and did not receive a DG from OTS. For this population, 10 CYOS instantaneous retention rate was 15.0% and at 20 CYOS, that rate was 2.6%.

Figure 17 shows the difference between the 60 unique survival lines and illustrates where the largest differences between the functions occur. Note that the gap between the highest and lowest retention rates among the different predictor combinations appears largest (approximately 20% difference in survival probability) around 4 CYOS and remains fairly constant until about 19 CYOS. A wide range of instantaneous retention rates at any given CYOS is a further indication that the attrition behavior is significantly different for NRO officers with different commissioning sources, genders, DG status, and enlisted experience.

Logistics (LOG).

The logistics (LOG) career field involves officers whose first digit in their core AFSC is a "2". LOG includes aircraft maintenance (21A), munitions and missile maintenance (21M) and logistics readiness (21R) officers. Stepwise regression revealed that prior enlisted service (5 categories), gender (2 categories), and distinguished graduate (2 categories) were all significant variables in determining the survival function. Addressing each combination of these settings resulted in 20 different survival functions. Figure 9 shows the survival function for non-prior enlisted LOG officers who did not receive distinguished graduate honors from their commissioning source. This plot includes a 95% confidence interval for the function.

Figure 9 reveals a steep decline in retention after 4 CYOS that levels out around the 10 CYOS point. Only 26.9% of the original LOG population illustrated is pro-
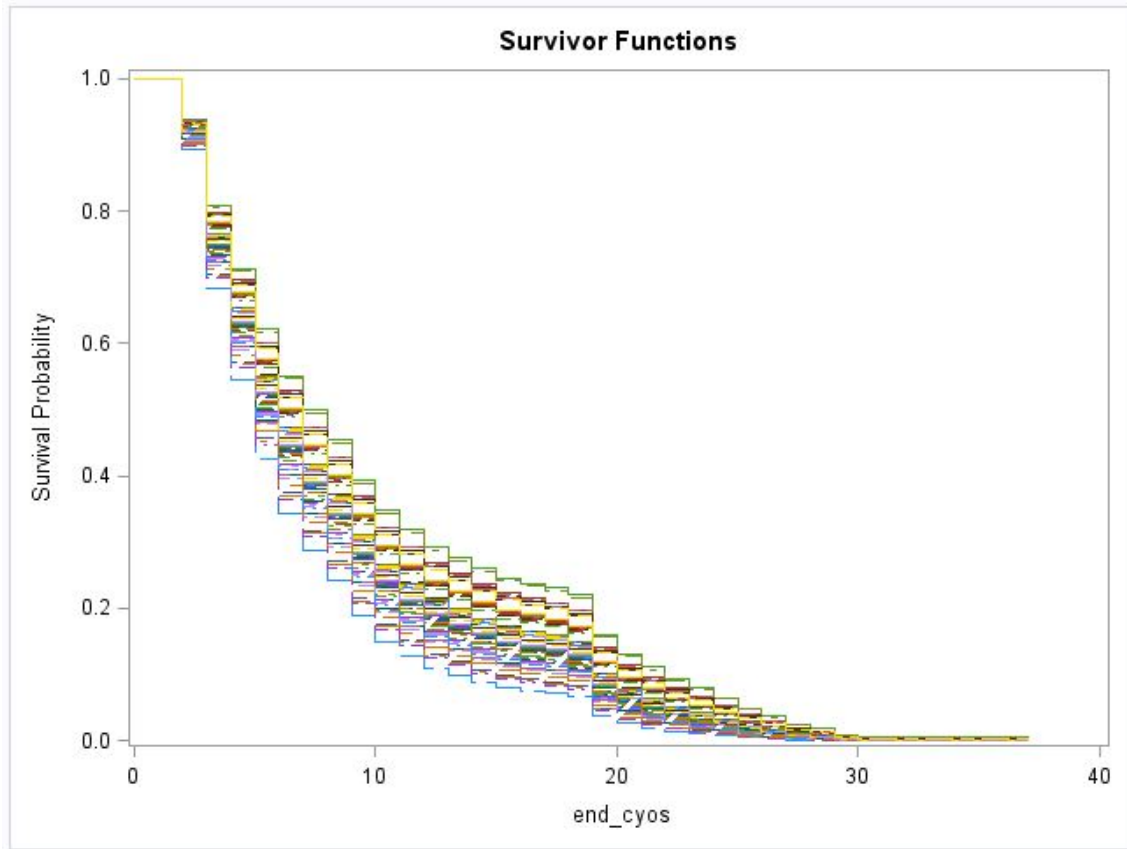
Figure 8. Non-Rated Operations Survival Functions for All 60 Variable Setting Combinations Illustrating Range of Functions
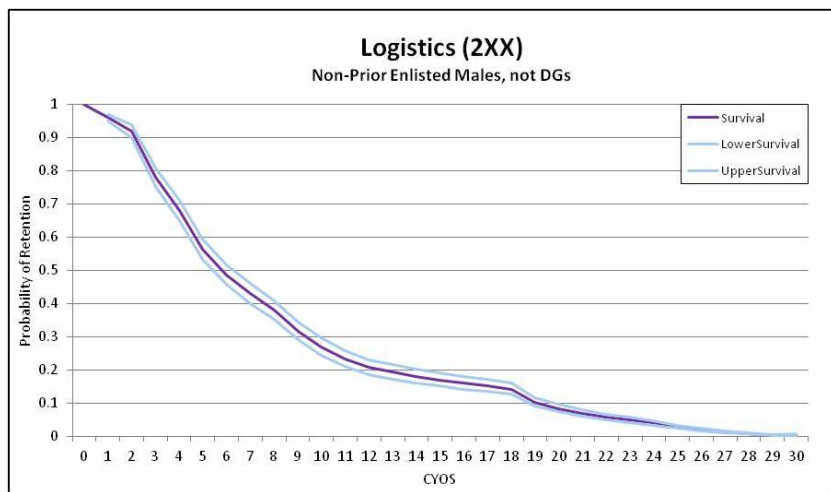
Figure 9. Logistics Survival Function (Non-Prior Enlisted Males, not DGs)

jected to still be in the Air Force at 10 CYOS and only 8.3% are expected to stay until 20 CYOS.

Of the 20 different survival functions for LOG officers, the best retention was found in males with more than 11 years of enlisted service who earned a DG from their commissioning source. At 10 CYOS, 36.1% are expected to still be in the Air Force and 14.6% are expected to remain at 20 CYOS. The worst retention of LOG officers was for females with 5-7 years of enlisted service and did not receive a DG. For this population, 10 CYOS instantaneous retention rate was 17.7% and at 20 CYOS, that rate was 3.8%.

Figure 10. Logistics Survival Functions for All 20 Variable Setting Combinations Illustrating Range of Functions

Figure 10 shows the difference between the 20 unique survival lines and illustrates where the largest differences between the functions occur. Note that the gap between the highest and lowest retention rates among the different predictor combinations

appears largest (approximately 12% difference in survival probability) around 4 CYOS and remains fairly constant until about 19 CYOS. A wide range of instantaneous retention rates at any given CYOS is a further indication that the attrition behavior is significantly different for LOG officers with different genders, DG status, and enlisted experience.

Support (SPT).

The support (SPT) career field involves officers whose first digit in their core AFSC is a "3". SPT includes security forces (31P), civil engineer (32E), communications and information (33S), band (35B), public affairs (35P), force support (38F), and personnel (36P) officers. Stepwise regression revealed that gender (2 categories) and distinguished graduate (2 categories) were significant variables in determining the survival function. Addressing each combination of these settings resulted in 4 different survival functions. Figure 11 shows the survival function for male SPT officers who did not receive distinguished graduate honors from their commissioning source. This plot includes a 95% confidence interval for the function.
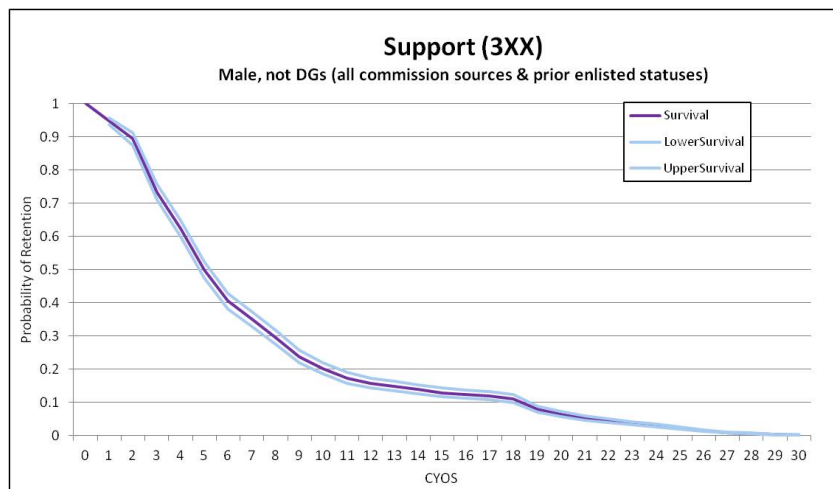
Figure 11. Support Survival Function (Male, not DGs)

Figure 11 reveals a steep decline in retention after 4 CYOS that levels out around the 10 CYOS point. Only 20.2% of the original SPT population illustrated is projected to still be in the Air Force at 10 CYOS and only 6.4% are expected to stay until 20 CYOS.

Of the 4 different survival functions for SPT officers, the best retention was found in males who earned a DG from their commissioning source. At 10 CYOS, 23.2% are expected to still be in the Air Force and 8.1% are expected to remain at 20 CYOS. The worst retention of SPT officers was for females who did not receive a DG from their respective commissioning sources. For this population, 10 CYOS instantaneous retention rate was 16.8% and at 20 CYOS, that rate was 4.7%.

Figure 12. Support Survival Functions for All 4 Variable Setting Combinations Illustrating Range of Functions

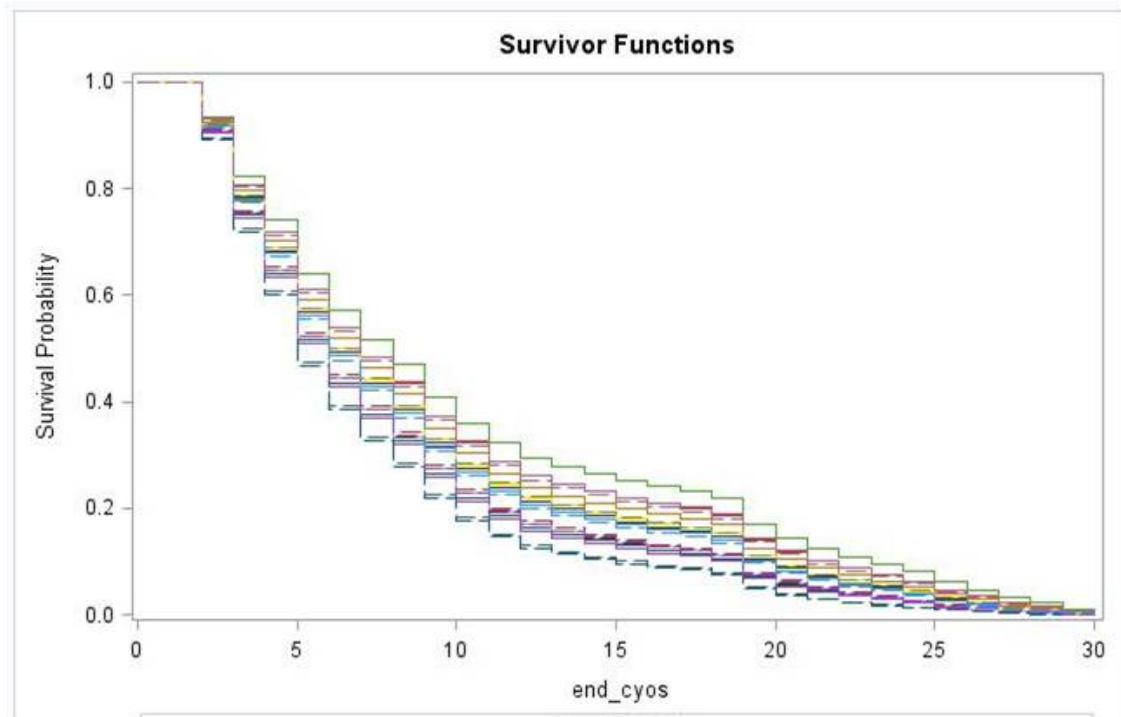Figure 12 shows the difference between the 4 unique survival lines and illustrates

where the largest differences between the functions occur. Note that the gap between the highest and lowest retention rates among the different predictor combinations appears largest (approximately 8% difference in survival probability) around 4 CYOS and remains fairly constant until about 12 CYOS. A wide range of instantaneous retention rates at any given CYOS is a further indication that the attrition behavior is significantly different for SPT officers with different genders and DG status.

Acquisitions (ACQ).

The acquisitions (ACQ) career field involves officers whose first digit in their core AFSC is a "6". ACQ includes operations research analyst (61A), behaviorial science/human scientist (61B), chemist (61C), physicist/nuclear engineer (61D), scientist (61S), developmental engineer (62E), acquisition manager (63A), contracting (64P), and financial management (65F) officers. Stepwise regression revealed that prior enlisted service (5 categories) and commissioning source (3 categories) were significant variables in determining the survival function. Addressing each combination of these settings resulted in 15 different survival functions. Figure 13 shows the survival function for ACQ officers with no prior enlisted experience and graduated from any one of the Service Academies. This plot includes a 95% confidence interval for the function.

Figure 13 reveals a steep decline in retention after 4 CYOS that levels out around the 10 CYOS point. Only 22.2% of the original ACQ population illustrated is projected to still be in the Air Force at 10 CYOS and only 5.1% are expected to stay until 20 CYOS.

Of the 15 different survival functions for ACQ officers, the best retention was found in ROTC graduates with more than 11 years of enlisted service. At 10 CYOS, 30.9% are expected to still be in the Air Force and 9.8% are expected to remain
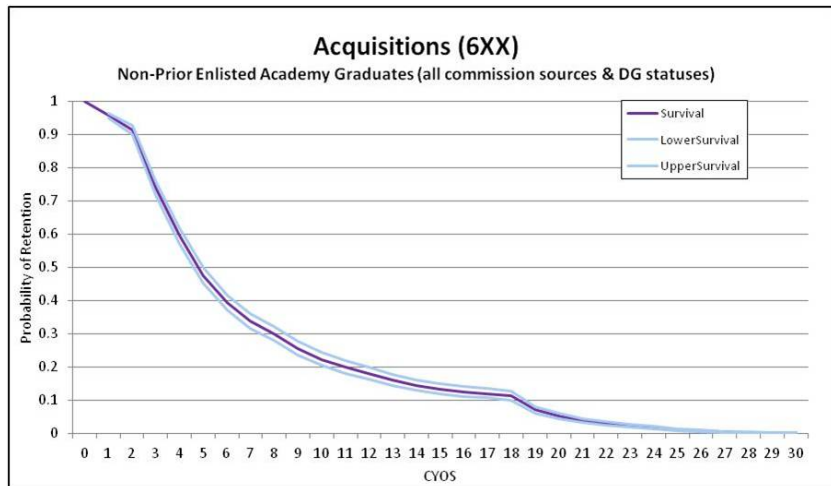
Figure 13. Acquisitions Survival Function (Non-Prior Enlisted Academy Graduates)

at 20 CYOS. The worst retention of ACQ officers was for Officer Training School (OTS) graduates with 8-11 years of enlisted service. For this population, 10 CYOS instantaneous retention rate was 18.6% and at 20 CYOS, that rate was 3.6%.

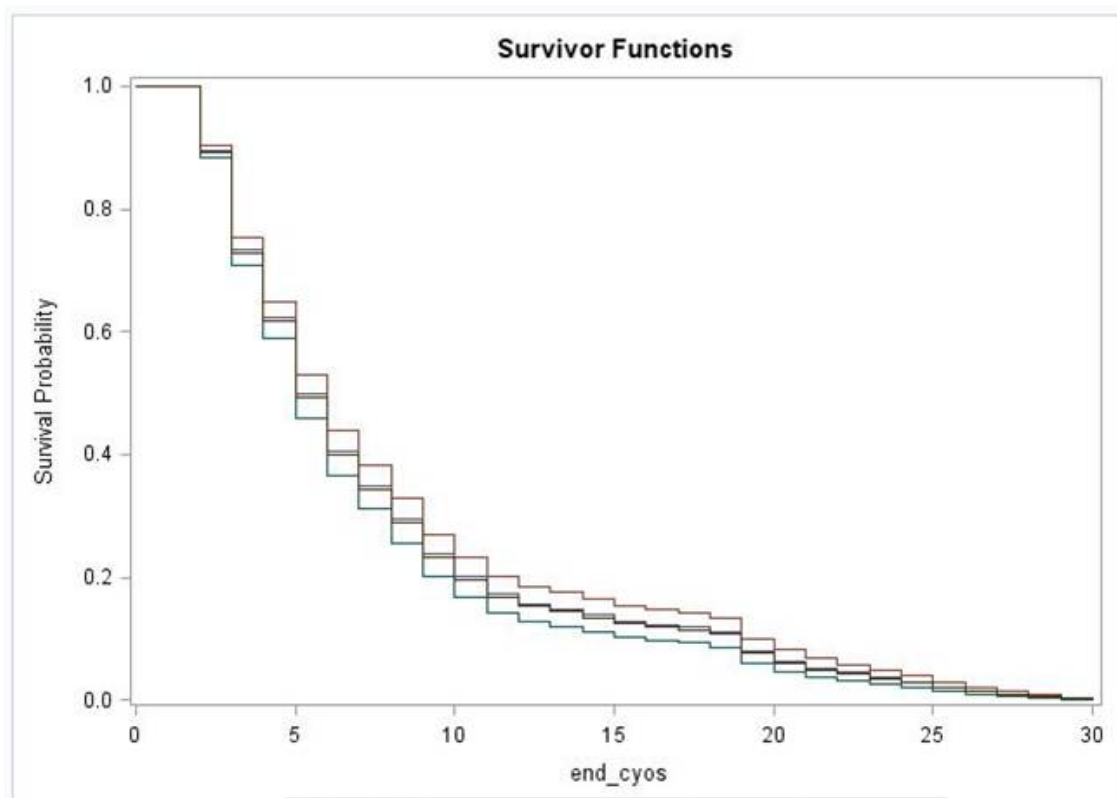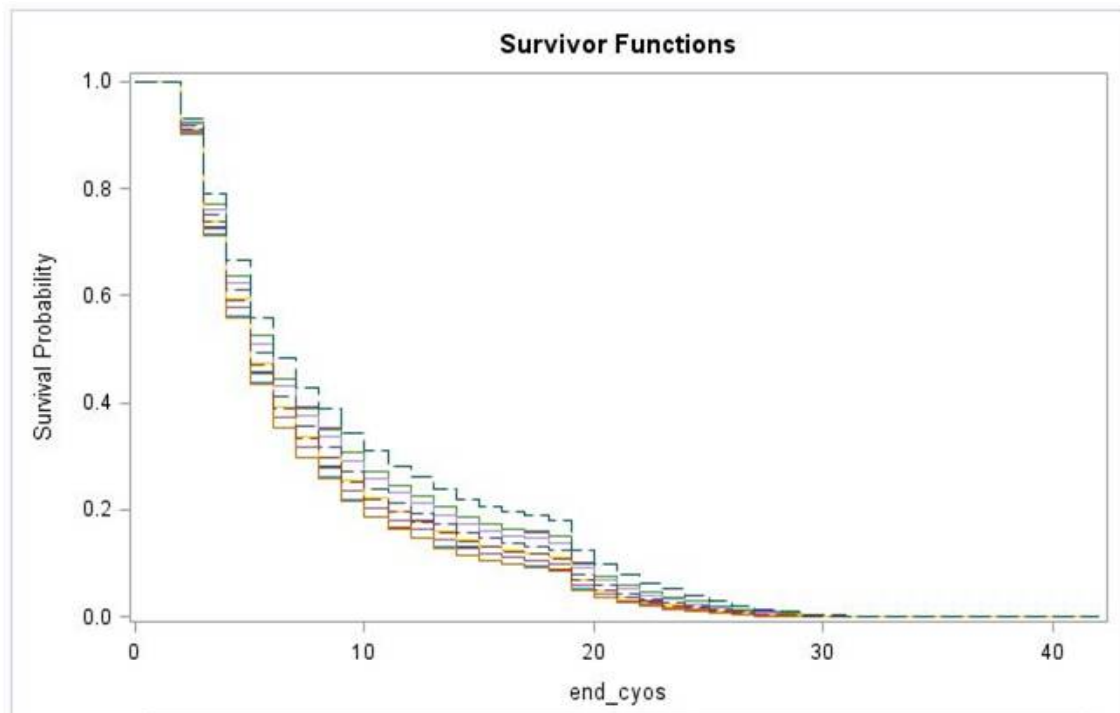**Figure 14. Acquisitions Survival Functions for All 15 Variable Setting Combinations Illustrating Range of Functions**

Figure 14 shows the difference between the 15 unique survival lines and illustrates where the largest differences between the functions occur. Note that the gap between the highest and lowest retention rates among the different predictor combinations appears largest (approximately 10% difference in survival probability) around 4 CYOS and remains fairly constant until about 18 CYOS. A wide range of instantaneous retention rates at any given CYOS is a further indication that the attrition behavior is significantly different for ACQ officers with different commissioning sources and enlisted experience.

Non-Rated Line (NRL).

The non-rated line (NRL) corps involves officers from the NRO, LOG, SPT, and ACQ career fields, as well as Office of Special Investigations (71S). Stepwise regression revealed that prior enlisted service (5 categories), gender (2 categories), commissioning source (3 categories), and distinguished graduate (2 categories) were all significant variables in determining the survival function. Addressing each combination of these settings resulted in 60 different survival functions. Figure 15 shows the survival function for NRL officers with no prior enlisted experience, who are male and graduated from any one of the Service Academies, but did not receive distinguished graduate honors from their commissioning source. This plot includes a 95% confidence interval for the function.

Figure 15 reveals a steep decline in retention after 4 CYOS that levels out around the 10 CYOS point. Only 22.8% of the original NRL population illustrated is projected to still be in the Air Force at 10 CYOS and only 6.1% are expected to stay until 20 CYOS.

Of the 60 different survival functions for NRL officers, the best retention was found in male ROTC graduates with more than 11 years of enlisted service who earned a
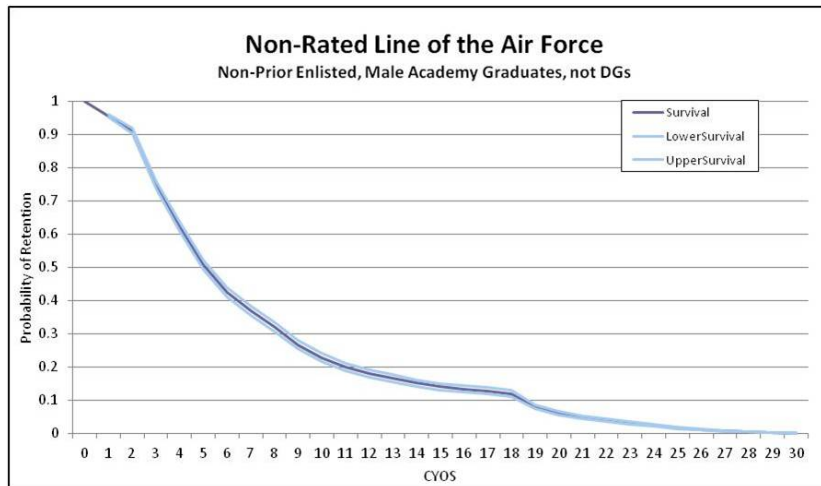
Figure 15. Non-Rated Line Survival Function (Non-Prior Enlisted, Male Academy Graduates, not DGs)

DG from their commissioning source. At 10 CYOS, 32.8% are expected to still be in the Air Force and 12.2% are expected to remain at 20 CYOS. The worst retention of NRL officers was for females who went to Officer Training School (OTS) with 3-4 years of enlisted service and did not receive a DG from OTS. For this population, 10 CYOS instantaneous retention rate was 16.2% and at 20 CYOS, that rate was 3.2%.

Figure 16 shows the difference between the 60 unique survival lines and illustrates where the largest differences between the functions occur. Note that the gap between the highest and lowest retention rates among the different predictor combinations appears largest (approximately 18% difference in survival probability) around 5 CYOS and remains fairly constant until about 14 CYOS. A wide range of instantaneous retention rates at any given CYOS is a further indication that the attrition behavior is significantly different for NRL officers with different commissioning sources, genders, DG status, and enlisted experience.
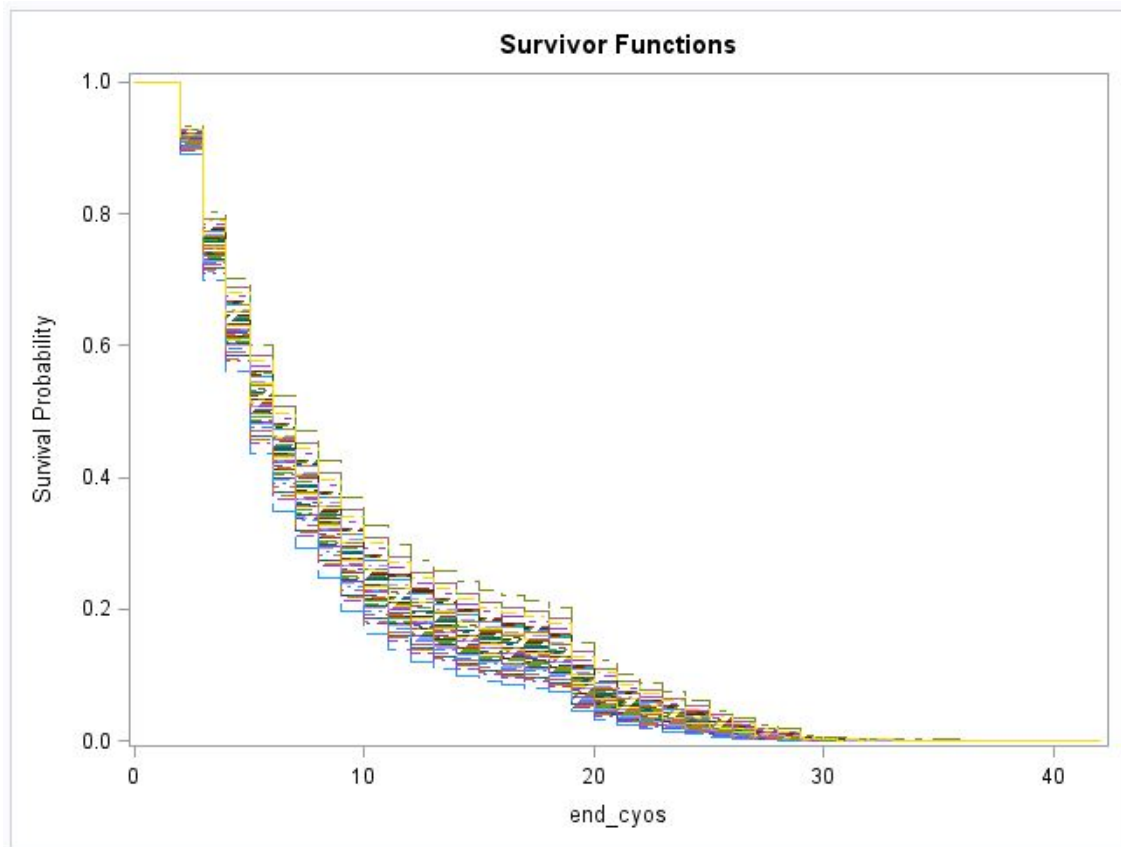
Figure 16. Non-Rated Line Officer Survival Functions for All 60 Variable Setting Combinations Illustrating Range of Functions

6.3 Application

Having up to 60 different survival lines for a single career field is not particularly useful for decision makers, since they are not interested in such a detailed level of fidelity. Typically, the attrition rates are broken out by AFSC, but in this study we reviewed only to the level of career fields. Due to its level of complexity, the example of how to apply the findings of this survival analysis is applied to the Non-Rated Operations (NRO) career field.

Although the survival analysis yielded 60 survival functions, some of them are not feasible or applicable. For example, officers who commission through the Academy must have fewer than 4 enlisted years of service; therefore, any combination of factors

that include "Academy" as the commissioning source and more than 4 years of prior service are not feasible. Additionally, only the top 10% of graduates earn a DG from their commissioning source, and if their commissioning source is Academy or ROTC, they get the first pick of AFSC. An overwhelming percentage of officers who earned a DG chose AFSCs in the rated or medical career fields, so many of the NRL survival curves that include a positive response for DG are not applicable.

One way to characterize a career field's attrition behavior over the next 30 years using survival analysis is to calculate using the newest officers in the Air Force. Reviewing the most recent end-of-year data, only those officers who have not completed their first CYOS (they are in 0 CYOS) were considered. These officers were then classified by career field, gender, commission source, prior enlisted years of service, and DG status using the same methodology as the previous data sets. A sum of officers for each variable setting combination was tabulated and the percentage of the career field calculated. Table 4 details the findings for NRO officers and illustrates that, while there are 60 different possible survival functions, only 20 are applicable to the newest NRO officers (0 CYOS in 2013).

To consolidate the 20 applicable NRO survival functions, each one must be weighted by the percentage of the population it represents. The percentages in 4 were multiplied by the respective survival functions they represent, then added together to provide a comprehensive look at how officers in the NRO career field will likely retain over the next 30 years. The resulting survival function is graphed in Figure 17. Figure 18 illustrates the difference between the consolidated NRO survival function and the aforementioned NRO survival function for non-prior enlisted, male Academy graduates who were not DGs (Figure 15). Both functions have basically the same shape, but the consolidated line has a lower retention rate than the specific population it is compared to from 4-25 CYOS.

Table 4. Non-Rated Operations Survival Functions Applicable to 0 CYOS in 2013.

Gender	Commission	Prior Enl	DG	Count	Percent
M	*Academy*	$0-2$	*No*	116	0.163380
M	*Academy*	$3-4$	*No*	4	0.005634
M	*OTS*	$0-2$	*No*	55	0.077465
M	*OTS*	$3-4$	*No*	14	0.019718
M	*OTS*	$5-7$	*No*	26	0.036620
M	*OTS*	$8-11$	*No*	40	0.056338
M	*OTS*	>11	*No*	32	0.045070
M	*ROTC*	$0-2$	*No*	217	0.305634
M	*ROTC*	$0-2$	*Yes*	6	0.008451
M	*ROTC*	$3-4$	*No*	6	0.008451
M	*ROTC*	$5-7$	*No*	5	0.007042
F	*Academy*	$0-2$	*No*	48	0.067606
F	*OTS*	$0-2$	*No*	25	0.035211
F	*OTS*	$0-2$	*Yes*	1	0.001408
F	*OTS*	$3-4$	*No*	3	0.004225
F	*OTS*	$5-7$	*No*	3	0.004225
F	*OTS*	$8-11$	*No*	4	0.005634
F	*OTS*	>11	*No*	12	0.016901
F	*ROTC*	$0-2$	*No*	92	0.129577
F	*ROTC*	$0-2$	*Yes*	1	0.001408
			TOTAL	710	1.00

6.4   Verification

To validate the model, HAF/A1PF provided the end-of-year 2014 data set. Comparing the personnel who were in the non-rated Line of the Air Force in 2013 with those who remained in 2014 provided empirical data to which the survival analysis model can be compared. Additionally, HAF/A1PF provided the output to their sustainment model, which was used to provide a reference for the expected accuracy of personnel models.

Accuracy was determined by averaging the percent error over all relevant factor combinations for each CYOS in each career field. Percent error was defined as the
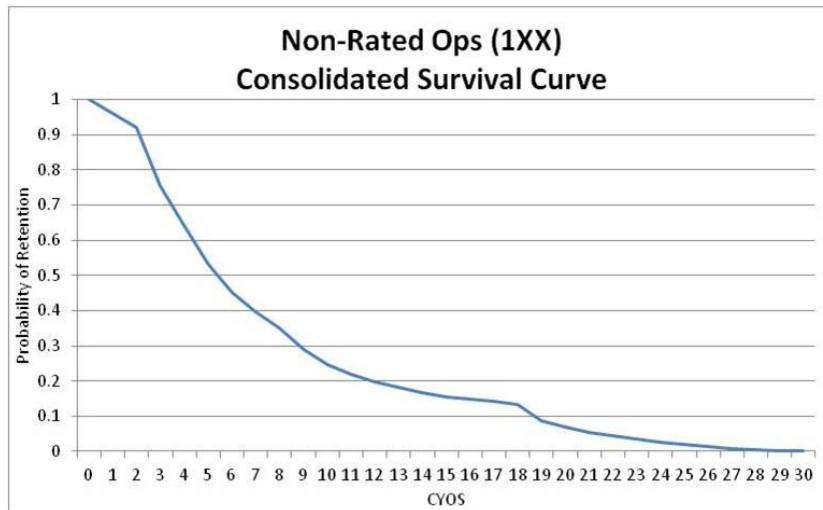
Figure 17. Non-Rated Operations Consolidated Survival Function
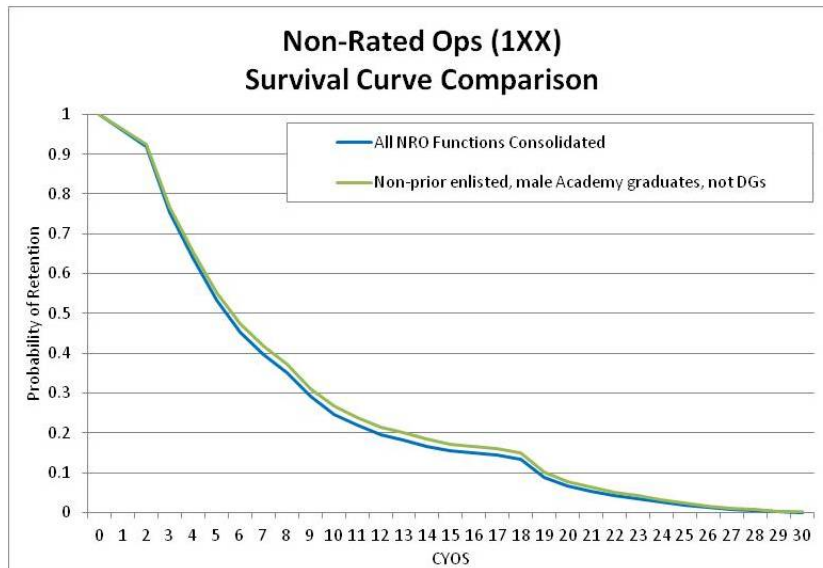
Figure 18. Non-Rated Operations Survival Function Comparison

percent difference between the observed retention rate and the retention rate forecast by the model. The survival analysis models divide personnel into smaller populations than used in the HAF/A1PF models, resulting in some bins having only one officer. If that officer retained, the one-year retention rate for that variable setting in that CYOS is 100% and if that officer attrits, the result is 0% one-year retention. None

54

of the survival models reflect a 100% or 0% one-year retention rate; therefore, the survival models will never be 100% accurate for a single year of data.

The HAF/A1PF model references specific AFSCs, not career fields, so the overall percent error for the "career field" was calculated by taking the average of the percent error for each of the entries for the AFSCs that were included in that same career field in the survival models. Table 5 illustrates the accuracy of the survival models, as well as the relative populations from the HAF/A1PF models. The survival models performed better than the HAF/A1PF models at predicting the 2014 retention of NRO and LOG officers, but worse with SPT and ACQ officers. Although variation in actual retention behavior is expected, further complicating the issue is the fact that 2014 was a "force shaping" year. The Air Force implemented incentives for voluntary attrition, as well as involuntary measures to decrease the officer population, those effects were felt towards the end of calendar year 2014 and the beginning of 2015.

Table 5. Model Accuracy Comparison: Survival Models vs HAF/A1PF Sustainment Model.

	Average Percent Error	
Career Field	Survival Models	HAF/A1PF Sustainment
NRO	15.2%	17.6%
LOG	16.0%	22.8%
SPT	23.2%	19.7%
ACQ	18.0%	12.4%

6.5 Summary

A total of 99 unique survival functions best characterize the data for use in attrition studies using survival analysis and proportional hazards regression. Each of the survival functions provides a more specific representation of historic behavior

that can be used to predict and/or shape future behavior. To best present the data to decision-makers, the unique survival functions must be combined, such as using weight according to the respective percentage of the populations they represent and aggregated. Survival analysis provides promising results when compared to the sustainment model currently used by HAF/A1PF when referring to the retention rates from calendar year 2013 to calendar year 2014.

VII.   Conclusion

7.1   Limitations of Work

This study was limited to characterizing the attrition behavior of various career fields in the non-rated Line of the Air Force. Typically, decision-makers are interested in how many personnel are needed in each yeargroup to meet mission requirements, as specified for each Core AFSC. The findings of this study are the first step to providing guidance for those decision-makers.

7.2   Follow-On Research

This study was meant as a baseline for a series of studies supporting Headquarters Air Force Directorate of Personnel (HAF/A1) and to create opportunities for further studies. Continuing along the lines of characterizing attrition (or retention) behavior, different methodologies could be used to explain historic trends and/or predict future behavior. Some methods to consider are simulation, network theory, stocks and flows, and linear programming. Forecasting methodologies such as Box Jenkins autoregressive integrated moving average (ARIMA) or time series decomposition may also provide insight to attrition behavior.

Another follow-on to this research could continue with survival analysis, but expand the scope. This study focused on non-rated Line of the Air Force officers, but future studies could include rated, medical, legal, and religious officers. Additionally, instead of examining career fields, a future study could investigate individual AFSCs.

Ultimately, it would be ideal to see how the aggregated attrition rates (survival functions) can be used to recommend force structure to HAF/A1 by way of a "new and improved" methodology. The current "sustainment lines" are used to develop accession targets, adjust ROTC scholarships and OTS throughput, and to "right

size" the force at the Core AFSC level. Although the current methodology has been accepted by the personnel management community at large, the details are difficult for the layperson to understand, so a more transparent, defendable model would enhance "buy-in" from decision makers.

7.3   Conclusion

Given the accuracy rates of the survival analysis compared to the rates embedded in the HAF/A1PF models, survival analysis has proven to be about as effective as the current model in characterizing the attrition rates of non-rated Line of the Air Force officers. This methodology should be considered as an alternative to the current models because of its accuracy and ease of explanation.

Appendix A. SAS Code for Logistic Regression

```
LIBNAME offdata ' ';  /* Where the "original" data and nrlaf cohort are stored */
LIBNAME stuff pcfiles path=' ';  /* This is the output file */
%let st_yr=1999; /* START year for data */
%let end_yr=2013; /* END year for data */
/* The create_cohort macro takes the LAF cohort and makes year-independent vars*/
/* such as gender, lastCF (career field), cmsn_src (commissioning source),     */
/* yrgrp (commissioned yeargroup), prior (1=yes, 2=no, 5=unk), enlYrs (number  */
/* of years enlisted--in bins), CF (career field - NRO=non-rated ops (1XX),    */
/* LOG=logistics (2XX), SPT=support (3XX), ACQ=acquisitions (6XX), OSI (7XX))  */
/* and commission (separates into OTS, Academy, ROTC, OTS DG, Academy DG, and  */
/* ROTC DG).     */
/*****************************************************************************/
/*                         ASSUMPTIONS      */
/*****************************************************************************/
/* 1. The most current non-empty record is the most accurate to determine   */
/*    gender, commissioning source, etc.       */
/* 2. The most current core AFSC is the only one that officer had over the  */
/*    entire career.  Since this cohort is only used for logit, this was OK */
/* 3. Prior enlisted have at least 3 yrs of enlistment. Sometimes ROTC dates */
/*    mess things up...and they don't get O-1E unless they have 3 years anyway. */
/* 4. The data files provided by HAF/A1 are accurate. Those files used to create */
/*    the nrlaf_cohort file was already "scrubbed" to fill in missing fields based */
/*    on other variables, to update AFSCsthat have merged/ceased to exist, fill in */
/*    Core AFSCs for rated officers, etc. The accuracy ofthe output data depends */
/*    on the accuracy of the macros used to create the input data. */
/* 5. If someone completes a CYOS, but separates before December, then they */
/*    do not get "credit" for completing that CYOS. */
/*****************************************************************************/
%macro create_cohort;
  data offdata.cohort_logreg;
    set offdata.nrlaf_cohort;
/* set up arrays for each variable to work backwards to find the most recent record*/
    array incl {&st_yr.:&end_yr.}  in_&st_yr.-in_&end_yr.;
    array Cyrs {&st_yr.:&end_yr.}  CYOS_&st_yr.-CYOS_&end_yr.;
    array sex  {&st_yr.:&end_yr.}  gender_&st_yr.-gender_&end_yr.;
    array core {&st_yr.:&end_yr.}  COREModel_&st_yr.-COREModel_&end_yr.;
    array comm {&st_yr.:&end_yr.}  commSrcLong_&st_yr.-commSrcLong_&end_yr.;
    array cyg  {&st_yr.:&end_yr.}  doc_&st_yr.-doc_&end_yr.;
    array enli {&st_yr.:&end_yr.}  priorYrs_&st_yr.-priorYrs_&end_yr.;
    array dos  {&st_yr.:&end_yr.}  DOS_&st_yr.-DOS_&end_yr.;
/* initiate each variable as blank.  This ensures each entry is the max length */
    gender = " ";
    lastCF = "   ";
    cmsn_src = "                         ";
    yrgrp = "    ";
    prior = 5;
    enlYrs = "   ";
    CF = "   ";
    commission="       ";
    sep=0;
/* work through cohort backwards to define year-independent variables */
    do yr = &end_yr. to &st_yr. by -1;
      If incl(yr) = 1 then do; /* only do if the officer was in the AF that year */
        If gender = " " then do;  /* fill in gender with most recent record */
          If sex(yr) = "M" then gender = "M";
          If sex(yr) = "F" then gender = "F";
        End;
        If lastCF = "   " and core(yr) ne "" then lastCF = core(yr);
          /* fill in lastCF with most recent record */
        If cmsn_src = "                         " and comm(yr) ne "" then do;
          /* fill in commission source with most recent record */
          cmsn_src = comm(yr);
          if cmsn_src in ("U.S.A.F. ACADEMY","U.S.NAVAL ACADEMY","US MILITARY
           ACAD","AFACDDG","OTHACDG")  then commission="Academy";
          else if cmsn_src in ("OCS GRADUATE","USAF OTS GRADUATE","DG OCS GRADUATE",
           "DG OTS GRADUATE") then commission="OTS    ";
          else commission="ROTC   ";
          if cmsn_src in ("AFACDDG","OTHACDG","DG OCS GRADUATE","DG OTS GRADUATE",
           "DG ROTC 2-YR PGM","DG ROTC 2-YR(FAG)", "DG ROTC 4-YR PGM",
           "DG ROTC 4-YR(FAG)") then dg=1;
          else dg=0;
        End;   /* fill in yrgrp with most recent record */
        If yrgrp = "    " and cyg(yr) ne "" then yrgrp = year(cyg(yr));
          /* fill in prior with most recent record */
        If prior = 5 and enli(yr) ne "" then do;
          If enli(yr) >= 3 then do;
            prior = 1;
            if enli(yr) <= 4 then enlYrs = "3-4";
            else if enli(yr) < 8 then enlYrs = "5-7";
            else if enli(yr) < 12 then enlYrs = "8-1";
            else enlYrs = ">11";
          End;
```

```
         Else do;
            prior = 0;
            enlYrs = "0-2";
          End;
        End;
        /*if officer has separate date, so use this date for CYOS in CCRs*/
        if dos(yr) ne "" then do;
          sepyrs=year(dos(yr))-yrgrp;
          sep=1;
        end;
      End;
    end;
    If prior = 5 then do; /*If enlisted years still empty , then not prior*/
      prior = 0;
      enlYrs = "0-2";
    end;
    If CF = "   " and lastCF ne "" then do; /* fill in CF with most recent record */
      Temp = substr(lastCF,1,1); /* First digit of Core AFSC determines CF */
      If temp = "1" then CF = "NRO";
       Else if temp = "2" then CF = "LOG";
       Else if temp = "3" then CF = "SPT";
       Else if temp = "6" then CF = "ACQ";
       Else if temp = "7" then CF = "OSI";
       Else CF = "OTH";
    End;
    /* work through cohort from oldest entries to newest to determine retention */
    do i=&st_yr. to (&end_yr. - 2);
      If Cyrs(i) ne "" then do;
        If Cyrs(i) = 0 and i < (&end_yr. - 5) then do;
          /* if you completed 0 years of service at the end of the first year and */
          /* there are still 5 more years of data in the cohort */
          If Cyrs(i+5) > 4 then CCR_0_6 = "Retain";
          /* if, 6 years later, you have more than 4 CYOS, then you retained */
          Else do;
            if sep=1 and sepyrs >= 6 then CCR_0_6 = "Retain";
            else CCR_0_6 = "Attrit";
          end;
          /* otherwise, you didn't finish 6 commissioned years of service */
        end;
        Else if Cyrs(i) = 4 and i< (&end_yr. - 3) then do;
          If Cyrs(i+4) > 6 then CCR_4_8 = "Retain";
          Else do;
            if sep=1 and sepyrs >= 8 then CCR_4_8 = "Retain";
            Else CCR_4_8 = "Attrit";
          end;
        end;
        Else if Cyrs(i) = 8 and i < (&end_yr. - 5) then do;
          If Cyrs(i+6) > 12 then CCR_8_14 = "Retain";
          else do;
            if sep=1 and sepyrs >= 12 then CCR_8_14 = "Retain";
            Else CCR_8_14 = "Attrit";
          end;
        end;
        Else if Cyrs(i) = 12 and i < (&end_yr. - 6) then do;
          If Cyrs(i+7) > 17 then CCR_12_19 = "Retain";
          else do;
            if sep=1 and sepyrs >= 19 then CCR_12_19 = "Retain";
            Else CCR_12_19 = "Attrit";
          end;
        end;
        Else if Cyrs(i) = 20 and i < (&end_yr. - 1) then do;
          If Cyrs(i+2) > 20 then CCR_20_22 = "Retain";
          else do;
            if sep=1 and sepyrs >= 22 then CCR_0_6 = "Retain";
            Else CCR_20_22 = "Attrit";
          end;
        end;
      End;
    End;
    if yrgrp ne "" and yrgrp < 3000;
  Run;
%mend;
%create_cohort;
/* The ccr_datasets macro makes separate datasets (temporary ones, stored in the*/
/* work folder) for each CCR to be used in the proc logistic section below.    */
/* They are excerpts from the offdata.cohort_logreg just created.  */
%macro ccr_datasets;
  data years_0_6 (keep=serial_id yrgrp gender commission enlYrs CF retain dg prior);
    set offdata.cohort_logreg
        (keep=serial_id yrgrp gender commission enlYrs CF ccr_0_6 dg prior
         where=(ccr_0_6 in ("Retain","Attrit")));
    if ccr_0_6="Retain" then retain=1;
    else retain=0;
  run;
  data years_4_8 (keep=serial_id yrgrp gender commission enlYrs CF retain dg prior);
    set offdata.cohort_logreg
        (keep=serial_id yrgrp gender commission enlYrs CF ccr_4_8 dg prior
         where=(ccr_4_8 in ("Retain","Attrit")));
    if ccr_4_8="Retain" then retain=1;
```

```
      else retain=0;
   run;
   data years_8_14 (keep=serial_id yrgrp gender commission enlYrs CF retain dg prior);
      set offdata.cohort_logreg
          (keep=serial_id yrgrp gender commission enlYrs CF ccr_8_14 dg prior
           where=(ccr_8_14 in ("Retain","Attrit")));
      if ccr_8_14="Retain" then retain=1;
      else retain=0;
   run;
   data years_12_19 (keep=serial_id yrgrp gender commission enlYrs CF retain dg prior);
      set offdata.cohort_logreg
          (keep=serial_id yrgrp gender commission enlYrs CF ccr_12_19 dg prior
           where=(ccr_12_19 in ("Retain","Attrit")));
      if ccr_12_19="Retain" then retain=1;
      else retain=0;
   run;
   data years_20_22 (keep=serial_id yrgrp gender commission enlYrs CF retain dg prior);
      set offdata.cohort_logreg
          (keep=serial_id yrgrp gender commission enlYrs CF ccr_20_22 dg prior
           where=(ccr_20_22 in ("Retain","Attrit")));
      if ccr_20_22="Retain" then retain=1;
      else retain=0;
   run;
%mend;
%ccr_datasets;
%macro log_regression;
   ods listing close;
   proc logistic data=years_0_6 descending;
      class yrgrp gender commission enlYrs CF dg prior;
      model retain = yrgrp gender commission enlYrs CF dg prior;
      ods output Type3=yrs06;
   run;
   proc export data=yrs06  outfile=stuff   DBMS=EXCELCS REPLACE;  run;
   proc logistic data=years_4_8 descending;
      class yrgrp gender commission enlYrs CF dg prior;
      model retain = yrgrp gender commission enlYrs CF dg prior;
      ods output Type3=yrs48;
   run;
   proc export data=yrs48  outfile=stuff  DBMS=EXCELCS REPLACE;  run;
   proc logistic data=years_8_14 descending;
      class yrgrp gender commission enlYrs CF dg prior;
      model retain = yrgrp gender commission enlYrs CF dg prior;
      ods output Type3=yrs814;
   run;
   proc export data=yrs814  outfile=stuff   DBMS=EXCELCS REPLACE;  run;
   proc logistic data=years_12_19 descending;
      class yrgrp gender commission enlYrs CF dg prior;
      model retain = yrgrp gender commission enlYrs CF dg prior;
      ods output Type3=yrs1219;
   run;
   proc export data=yrs1219  outfile=stuff  DBMS=EXCELCS REPLACE;  run;
   proc logistic data=years_20_22 descending;
      class yrgrp gender commission enlYrs CF dg prior;
      model retain = yrgrp gender commission enlYrs CF dg prior;
      ods output Type3=yrs2022;
   run;
   proc export data=yrs2022  outfile=stuff  DBMS=EXCELCS REPLACE;  run;
%mend;
%log_regression;
```

Appendix B. SAS Code for Survival Analysis

```
LIBNAME offhist ' '; /* input folder */
LIBNAME offdata ' '; /* output folder */
LIBNAME survive pcfiles path=' ';  /* This is the output file */
%let st_yr=1999; /*first year of data to run*/
%let end_yr=2013;/*last year of data to run*/
%let afsclist=NRO LOG SPT ACQ OSI; /* CF list not including NRL */
%let afsclist2=NRL NRO LOG SPT ACQ OSI; /* CF list including NRL (entire dataset) */
option spool;
/* The code is organized in macros for easier reading and running */
%macro ScopeData;
  %do yr=&st_yr. %to &end_yr.;
      /*Create a new dataset with only the variables and officers needed*/
    data offdata.sortinv&yr.;
      set offhist.eoyinv&yr. (keep=serial_id aaw ahb aqf aqt aem300 CYOS_EOP YOS_EOP
            asb6 ase PS_EOP DOC grade Source_of_Commission COREModel afc207
            where=(aaw not in ('B30','B31') and ahb ne 'X' and aqf le '39' and
            aqt ne '3' and aem300='A' and COREModel not in ('13A','13B') and
            substr(COREModel,1,2) not in ('11','12','18','51','ZZ','MS','MC','NC')
            and year(DOC) < 3000));
    run;
    /*This step sorts the data by the serial_id.*/
    proc sort data=offdata.sortinv&yr. out=offdata.sortinv&yr.;
      by serial_id;
    run;
    /*Get rid of filter variables then rename the variables that remain. */
    data offdata.sortinv&yr.;
      set offdata.sortinv&yr. (keep=serial_id CYOS_EOP asb6 PS_EOP DOC COREModel
                                 afc207 Source_of_Commission
          rename=(CYOS_EOP=CYOS_&yr COREModel=COREModel_&yr DOC=doc_&yr
            asb6=gender_&yr PS_EOP=priorYrs_&yr afc207=DOS_&yr
            Source_of_Commission=commSrcLong_&yr ));
      in_&yr=1;
    run;
  %end;
%mend;
%ScopeData;
%macro QuickCohort;
  /*This step just merges all of the datasets together by serial_id*/
  data offdata.NRlaf_surv_cohort;
    merge
      %do yr=&st_yr. %to &end_yr.;
        offdata.sortinv&yr.
      %end;;
      by serial_id;
  run;
%mend;
%QuickCohort;
%macro SurviveData;
  /* This macro creates the large dataset for the survival analysis code.*/
  data offdata.survive (keep=start_cyos end_cyos censor_af censor_cf yrgrp gender
          commission enlYrs cf dg afsc retain_af retain_cf);
    set offdata.NRlaf_surv_cohort;
    /* set up arrays for each variable */
    array incl {&st_yr.:&end_yr.}  in_&st_yr.-in_&end_yr.;
    array Cyrs {&st_yr.:&end_yr.}  CYOS_&st_yr.-CYOS_&end_yr.;
    array sex  {&st_yr.:&end_yr.}  gender_&st_yr-gender_&end_yr.;
    array core {&st_yr.:&end_yr.}  COREModel_&st_yr.-COREModel_&end_yr.;
    array comm {&st_yr.:&end_yr.}  commSrcLong_&st_yr.-commSrcLong_&end_yr.;
    array cyg  {&st_yr.:&end_yr.}  doc_&st_yr.-doc_&end_yr.;
    array enli {&st_yr.:&end_yr.}  priorYrs_&st_yr.-priorYrs_&end_yr.;
    array dos  {&st_yr.:&end_yr.}  DOS_&st_yr.-DOS_&end_yr.;
    length gender $1 lastCF $3 cmsn_src $30 yrgrp $4 enlYrs $3 CF $3 commission $7
            afsc $3;
    prior = 5;
    /* work through cohort backwards to define year-independent vars */
    /* the assumption is that the most recent record is the most accurate */
    Do yr = &end_yr. to &st_yr. by -1;
      If incl(yr) = 1 then do; /* only run if the officer was in the AF that year */
      /* GENDER */
        If gender = " " then do;
          If sex(yr) = "M" then gender = "M";
          If sex(yr) = "F" then gender = "F";
        End;
      /* COMMISSION SOURCE */
        If cmsn_src = "                              " and comm(yr) ne "" then do;
          cmsn_src = comm(yr);
          if cmsn_src in ("U.S.A.F. ACADEMY","U.S.NAVAL ACADEMY","US MILITARY ACAD",
              "AFACDDG","OTHACDG")  then commission="Academy";
          else if cmsn_src in ("OCS GRADUATE","USAF OTS GRADUATE","DG OCS GRADUATE",
              "DG OTS GRADUATE") then commission="OTS    ";
          else commission="ROTC   ";
```

```
         if substr(cmsn_src,1,2)="DG" or cmsn_src="AFACDDG" then dg=1;
         else dg=0;
     End;
   /* COMMISSIONED YEARGROUP */
     If yrgrp = "     " and cyg(yr) ne "" then yrgrp = year(cyg(yr));
   /* PRIOR YEARS OF SERVICE */
     If prior = 5 and enli(yr) ne "" then do;
       If enli(yr) >= 3 then do;
         prior = 1;
         if enli(yr) <= 4 then enlYrs = "3-4";
         else if enli(yr) < 8 then enlYrs = "5-7";
         else if enli(yr) < 12 then enlYrs = "8-1";
         else enlYrs = ">11";
       End;
       Else do;
         prior = 0;
         enlYrs = "0-2";
       End;
     End;
   /* CAREER FIELD */
     If CF = "     " and core(yr) ne "" then do;
       Temp = substr(core(yr),1,1); /* First digit of Core AFSC tells CF */
       If temp = "1" then CF = "NRO";
       Else if temp = "2" then CF = "LOG";
       Else if temp = "3" then CF = "SPT";
       Else if temp = "6" then CF = "ACQ";
       Else if temp = "7" then CF = "OSI";
       Else CF = "OTH";
     End;
   End;
End;
/*If enlisted years hasn't been filled in, then assume not prior*/
If prior = 5 then do;
   prior = 0;
   enlYrs = "0-2";
end;
/* This loop separates the data into single entries */
already_gone = 0;
do i=&st_yr. to (&end_yr. - 1);
   /* If the officer is in the AF during year i, then start a record */
   If incl(i) = 1 then do;
     start_cyos = Cyrs(i);
     afsc = core(i);
/* See if/when the officer separates or changes AFSC */
     do j=i+1 to (&end_yr. - 1);
       if incl(j) = 0 or core(j) ne afsc then do; /* changed AFSC or left service */
         if incl(j) = 0 then do; /* if left AF, assign ending values & output */
           end_cyos = Cyrs(j-1);
           censor_af = 1; /* 1-sep/retired, 0-still in at the end of data */
           censor_cf = 1; /* 1-sep/retired/crossflow, 0-still in */
           retain_af = 0; /* left AF */
           retain_cf = 0; /* left AFSC (career field) */
           OUTPUT;
           i = &end_yr.; /* since out of AF, no need to look at future */
           j = &end_yr.; /* records */
           already_gone = 1; /* left the AF.  This is for the last year of */
         end;                           /* data, which needs special treatment */
         else if core(j) ne afsc then do; /* if still in AF, but left CF, then */
                       /* assign ending values for the record then output */
           end_cyos = Cyrs(j-1);
           censor_af = 0; /* 1-seperated/retired, 0-still in at the end of data */
           censor_cf = 1; /* 1-seperated/retired/crossflow, 0-still in */
           retain_af = 1;  /* stayed in AF */
           retain_cf = 0;  /* left AFSC (career field) */
           OUTPUT;
           i = j-1; /* we have already reviewed the records up to year j, so */
                      /* increment i to look at the next record */
           j = &end_yr.;
           end;
         end;
       end;
     end;
   end;
   if already_gone = 0 then do; /* haven't already left AF */
     if incl(i) = 1 then do; /* still in AF at &endyr data set */
       if incl(i-1) = 1 then do; /* was in AF last year */
         if afsc = core(i) then do; /* stayed in same AFSC, so data "censored" */
           end_cyos = Cyrs(i);
           censor_af = 0; /* 1-seperated/retired, 0-still in at the end of data */
           censor_cf = 0; /* 1-sep/retire/crossflow, 0-still in */
           retain_af = 1;
           retain_cf = 1;
           OUTPUT;
         end;
         else do; /* changed AFSCs, so data is "censored" */
           end_cyos = Cyrs(i-1); /* changed AFSC, but stayed in AF.*/
           censor_af = 0; /* 1-seperated/retired, 0-still in at the end of data */
           censor_cf = 1; /* 1-sep/retire/crossflow, 0-still in */
           retain_af = 1;
```

```
            retain_cf = 0;
            OUTPUT;
          end;
        end;
      else delete; /* if is in AF during &endyr, but not previous year, this */
    end;              /* is 0 CYOS and won't add any info to model */
      else do; /* if not in AF at end of the last data set */
        if incl(i-1) = 1 then do; /* was in AF last year */
          end_cyos = Cyrs(i-1);
          censor_af = 1; /* 1-seperated/retired, 0-still in at the end of data */
          censor_cf = 1; /* 1-sep/retire/crossflow, 0-still in at the end of data */
          retain_af = 0;
          retain_cf = 0;
          OUTPUT;
        end;
      end;
    end;
  run;
  /* Make sure that the fields are all populated and "bum data" is not included */
  data offdata.survive;
    set offdata.survive;
    if start_cyos ne 0 and end_cyos ne 0;
    if gender ne " " and commission ne "         " and enlYrs ne "   " and
            afsc ne "    " and dg ne " " and afsc not in ("52R" "61X");
  run;
%mend;
%SurviveData;
/* The following macro allows calculation automation. Instead of performing a */
/* procedure on each CF and writing it out, we can use DO_OVER to only write */
/* the proc once, but perform it on every CF in the array (located at the very top */
/* of this code).  Code source is included. */
/*http://www.sascommunity.org/wiki/Tight_Looping_with_Macro_Arrays */
%MACRO DO_OVER(arraypos, array=, values=, delim=%STR( ), phrase=?, escape=?,
              between=, macro=, keyword=);
 /*  Last modified: 8/4/2006
Function: Loop over one or more arrays of macro variables substituting values
          into a phrase or macro.
Authors: Ted Clay, M.S.  (Clay Software & Statistics tclay@ashlandhome.net
         (541) 482-6435)
         David Katz, M.S.  (www.davidkatzconsulting.com)
"Please keep, use and pass on the ARRAY and DO_OVER macros with this authorship
         note.  -Thanks "
Send any improvements, fixes or comments to Ted Clay.
 Full documentation with examples appears in "Tight Looping with Macro Arrays".
 SUGI Proceedings 2006, found at http://www2.sas.com/proceedings/sugi31/040-31.pdf
 The keyword parameter was added after the SUGI article was written. */
 %LOCAL _IntrnlN
 _Intrnl1 _Intrnl2 _Intrnl3 _Intrnl4 _Intrnl5 _Intrnl6 _Intrnl7 _Intrnl8
 _Intrnl11 _Intrnl12 _Intrnl13 _Intrnl14 _Intrnl15 _Intrnl16 _Intrnl17 _Intrnl18
 _Intrnl21 _Intrnl22 _Intrnl23 _Intrnl24 _Intrnl25 _Intrnl26 _Intrnl27 _Intrnl28
 _Intrnl31 _Intrnl32 _Intrnl33 _Intrnl34 _Intrnl35 _Intrnl36 _Intrnl37 _Intrnl38
 _Intrnl41 _Intrnl42 _Intrnl43 _Intrnl44 _Intrnl45 _Intrnl46 _Intrnl47 _Intrnl48
 _Intrnl51 _Intrnl52 _Intrnl53 _Intrnl54 _Intrnl55 _Intrnl56 _Intrnl57 _Intrnl58
 _Intrnl61 _Intrnl62 _Intrnl63 _Intrnl64 _Intrnl65 _Intrnl66 _Intrnl67 _Intrnl68
 _Intrnl71 _Intrnl72 _Intrnl73 _Intrnl74 _Intrnl75 _Intrnl76 _Intrnl77 _Intrnl78
 _Intrnl81 _Intrnl82 _Intrnl83 _Intrnl84 _Intrnl85 _Intrnl86 _Intrnl87 _Intrnl88
_Intrnl91 _Intrnl92 _Intrnl93 _Intrnl94 _Intrnl95 _Intrnl96 _Intrnl97 _Intrnl98
 _Intrnl9 _Intrnl19 _Intrnl29 _Intrnl39 _Intrnl49 _Intrnl59 _Intrnl69 _Intrnl79
_Intrnl89_Intrnl99 _Intrnl90 _Intrnl100
_Intrnl10 _Intrnl20  _Intrnl30 _Intrnl40 _Intrnl50 _Intrnl60_Intrnl70_Intrnl80
 _KEYWRDN _KEYWRD1 _KEYWRD2 _KEYWRD3 _KEYWRD4 _KEYWRD5 _KEYWRD6 _KEYWRD7
 _KEYWRD8 _KEYWRD9 _KWRDI SOMETHINGTODO TP VAL VALUESGIVEN
 ARRAYNOTFOUND CRC CURRPREFIX DELIMI DID FRC I ITER J KWRDINDEX MANUM PREFIXES
 PREFIXN PREFIX1 PREFIX2 PREFIX3 PREFIX4 PREFIX5 PREFIX6 PREFIX7 PREFIX8 PREFIX9;
%let somethingtodo=Y;
%* Get macro array name(s) from either keyword or positional parameter;
%if       %str(&arraypos) ne %then %let prefixes=&arraypos;
%else %if %str(&array)    ne %then %let prefixes=&array;
%else %if %quote(&values) ne %then %let prefixes=_Intrnl;
%else %let Somethingtodo=N;
%if &somethingtodo=Y %then %do;
%* Parse the macro array names;
%let PREFIXN=0;
%do MAnum = 1 %to 999;
 %let prefix&MANUM=%scan(&prefixes,&MAnum,' ');
 %if &&prefix&MAnum ne %then %let PREFIXN=&MAnum;
 %else %goto out1;
%end;
%out1:
%* Parse the keywords;
%let _KEYWRDN=0;
%do _KWRDI = 1 %to 999;
 %let _KEYWRD&_KWRDI=%scan(&KEYWORD,&_KWRDI,' ');
 %if &&_KEYWRD&_KWRDI ne %then %let _KEYWRDN=&_KWRDI;
 %else %goto out2;
%end;
%out2:
%* Load the VALUES into macro array 1 (only one is permitted);
%if %length(%str(&VALUES)) >0 %then %let VALUESGIVEN=1;
```

```
%else %let VALUESGIVEN=0;
%if &VALUESGIVEN=1 %THEN %do;
  %* Check for numbered list of form xxx-xxx and expand it using NUMLIST macro.;
   %IF (%INDEX(%STR(&VALUES),-) GT 0) and (%SCAN(%str(&VALUES),2,-) NE ) and
       (%SCAN(%str(&VALUES),3,-) EQ ) %THEN %LET VALUES=%NUMLIST(&VALUES);
     %do iter=1 %TO 9999;
       %let val=%scan(%str(&VALUES),&iter,%str(&DELIM));
       %if %quote(&VAL) ne %then %do;
          %let &PREFIX1&ITER=&VAL;
          %let &PREFIX1.N=&ITER;
       %end;
     %else %goto out3;
   %end;
   %out3:
%end;
%let ArrayNotFound=0;
%do j=1 %to &PREFIXN;
   %*put prefix &j is &&prefix&j;
   %LET did=%sysfunc(open(sashelp.vmacro (where=(name eq "%upcase(&&PREFIX&J..N)"))));
     %LET frc=%sysfunc(fetchobs(&did,1));
     %LET crc=%sysfunc(close(&did));
     %IF &FRC ne 0 %then %do;
       %PUT Macro Array with Prefix &&PREFIX&J does not exist;
       %let ArrayNotFound=1;
   %end;
%end;
%if &ArrayNotFound=0 %then %do;
   %if %quote(%upcase(&BETWEEN))=COMMA %then %let BETWEEN=%str(,);
   %if %length(%str(&MACRO)) ne 0 %then %do;
     %let TP = %nrstr(%&MACRO)(;
       %do J=1 %to &PREFIXN;
         %let currprefix=&&prefix&J;
         %IF &J>1 %then %let TP=&TP%str(,);
* Write out macro keywords followed by equals. If fewer keywords than macro
         arrays, assume parameter is positional and do not write keyword=;
         %let kwrdindex=%eval(&_KEYWRDN-&PREFIXN+&J);
         %IF &KWRDINDEX>0 %then %let TP=&TP&&_KEYWRD&KWRDINDEX=;
         %LET TP=&TP%nrstr(&&)&currprefix%nrstr(&I);
       %END;
       %let TP=&TP);  %* close parenthesis on external macro call;
   %end;
   %else %do;
     %let TP=&PHRASE;
     %let TP = %qsysfunc(tranwrd(&TP,&ESCAPE._I_,%nrstr(&I.)));
     %let TP = %qsysfunc(tranwrd(&TP,&ESCAPE._i_,%nrstr(&I.)));
     %do J=1 %to &PREFIXN;
       %let currprefix=&&prefix&J;
       %LET TP = %qsysfunc(tranwrd(&TP,&ESCAPE&currprefix,
                          %nrstr(&&)&currprefix%nrstr(&I..)));
       %if &PREFIXN=1 %then %let TP = %qsysfunc(tranwrd(&TP,&ESCAPE,
                          %nrstr(&&)&currprefix%nrstr(&I..)));
     %end;
   %end;
   * resolve TP (the translated phrase) and perform the looping;
   %do I=1 %to &&&prefix1.n;
     %if &I>1 and %length(%str(&between))>0 %then &BETWEEN;
     %unquote(&TP)
   %end;
   %end;
%end;
%MEND;
%macro SurvivalCalc; /* Creates survival fns for each CF, and NRLAF as a whole. */
/* Analysis is performed for each CF, the data needs to be sorted accordingly */
  proc sort data=offdata.survive out=offdata.survive; by cf; run;
  ods graphics on;
/* Perform Survival Analysis on each AFSC */
  proc phreg data=offdata.survive outest=offdata.CF_coefficients
    plots(overlay)=survival;
    output out=offdata.survival_residuals /* this is the output dataset */
    lmax=sensitivity /* sensitivity of model to each observation */
    resdev=deviance_residual /* residuals to test model adequacy */
    resmart=martingale_residual /* residuals to test model adequacy */
    ressco=score_residual /* residuals to assess leverage by each individual */
    ressch=schoenfeld_residual /* residuals to check assumptions */
    xbeta=lin_pred_scores; /* used to plot against residuals to check lack of fit */
    by cf; /* This creates a different survival line for each afsc */
    class gender commission enlYrs dg; /* List categorical vars */
    model end_cyos*censor_cf(0) = gender commission enlYrs dg/entry=start_cyos
    /* list all vars after the = */ /* entry = designates the left-truncated data */
    selection=stepwise slentry=0.2 slstay=0.06;
    /* stepwise selection of vars for model: has to be significant at 0.2 level */
    /* to enter model, has to be significant at 0.06 level to stay */
  run;
/* Perform Survival Analysis on all officers in data set (consolidated) */
  proc phreg data=offdata.survive outest=offdata.AF_coefficients plots=(survival);
    output out=offdata.AFsurvival_residuals /* this is the output dataset */
    lmax=AFsensitivity /* sensitivity of model to each observation */
    resdev=AFdeviance_residual /* residuals to test model adequacy */
    resmart=AFmartingale_residual /* residuals to test model adequacy */
```

```
      ressco=AFscore_residual /* residuals to assess leverage by each individual */
      ressch=schoenfeld_residual /* residuals to assess assumptions */
      xbeta=lin_pred_scores; /* used to plot against residuals to check lack of fit */
      class gender commission enlYrs dg; /* List of categorical vars */
      model end_cyos*censor_cf(0) = gender commission enlYrs dg/entry=start_cyos
      /* list all vars after the = */  /* entry= designates the left-truncated data */
      selection=stepwise slentry=0.2 slstay=0.06;
       /* stepwise selection of vars for model: has to be signifcant at 0.2 level */
       /* to enter model, has to be significant at 0.06 level to stay */
    run;
  ods graphics off;
/* Combine the datasets. First, give CF a value in the consolidated cohort */
  data offdata.af_coefficients;
    set offdata.af_coefficients;
    cf="NRL"; /* NRL = non-rated line (all officers in the original cohort) */
  run;
  data offdata.all_coefficients (keep=cf genderF commissionAcademy commissionOTS
                     enlYrsON2 enlYrs3N4 enlYrs5N7 enlYrs8N1 dg0);
    set offdata.cf_coefficients
        offdata.af_coefficients;
  run;
%mend;
%SurvivalCalc;
%macro survival_functions(core_afsc);
/* This macro will create the cov datasets & use them to create survival fns */
/* This code looks at the stepwise regression above in the proc phreg sections */
/* and outputs the coefficients for the relevant var settings.  Most importantly, */
/* this returns the list of vars to be used to create the actual survival */
/* functions in later macros (16 combinations)  */
  data offdata.coef_surv_&core_AFSC. (keep=gender commission enlyrs dg reflist);
    set offdata.all_coefficients (where=(cf="&core_AFSC."));
    array sex  (2) $ sex1-sex2   ('M' 'F');
    array comm (3) $ comm1-comm3 ('Academy' 'OTS' 'ROTC');
    array enli (5) $ enli1-enli5 ('0-2' '3-4' '5-7' '8-1' '>11');
    if genderF > 0 then do MorF = 1 to 2;
      gender = sex(MorF);
      if commissionOTS > 0 then do cmsn = 1 to 3;
        commission = comm(cmsn);
        if enlYrsON2 > 0 then do sted = 1 to 5;
          enlYrs = enli(sted);
          if dg0 > 0 then do grad = 0 to 1;
            dg = grad;
            reflist = "gender commission enlYrs dg";
            OUTPUT; /* gender, commission, enlYrs, dg (all) signficant */
          end;
          else do;
            reflist = "gender commission enlYrs";
            OUTPUT; /* gender, commission, enlYrs signficant */
          end;
        end;
        else do;
          if dg0 > 0 then do grad = 0 to 1;
            dg = grad;
            reflist = "gender commission dg";
            OUTPUT; /* gender, commission, dg signficant */
          end;
          else do;
            reflist = "gender commission";
            OUTPUT; /* gender, commission signficant */
          end;
        end;
      end;
      else do;
        if enlYrsON2 > 0 then do sted = 1 to 5;
          enlYrs = enli(sted);
          if dg0 > 0 then do grad = 0 to 1;
            dg = grad;
            reflist = "gender enlYrs dg";
            OUTPUT; /* gender, enlYrs, dg signficant */
          end;
          else do;
            reflist = "gender enlYrs";
            OUTPUT; /* gender, enlYrs signficant */
          end;
        end;
        else do;
          if dg0 > 0 then do grad = 0 to 1;
            dg = grad;
            reflist = "gender dg";
            OUTPUT; /* gender, dg signficant */
          end;
          else do;
            reflist = "gender";
            OUTPUT; /* gender signficant */
          end;
        end;
      end;
    end;
    else do;
```

```
      if commissionOTS > 0 then do cmsn = 1 to 3;
        commission = comm(cmsn);
          if enlYrsON2 > 0 then do sted = 1 to 5;
            enlYrs = enli(sted);
            if dg0 > 0 then do grad = 0 to 1;
              dg = grad;
              reflist = "commission enlYrs dg";
              OUTPUT; /* commission, enlYrs, dg signficant */
            end;
            else do;
              reflist = "commission enlYrs";
              OUTPUT; /* commission, enlYrs signficant */
            end;
          end;
          else do;
            if dg0 > 0 then do grad = 0 to 1;
              dg = grad;
              reflist = "commission dg";
              OUTPUT; /* commission, dg signficant */
            end;
            else do;
              reflist = "commission";
              OUTPUT; /* commission signficant */
            end;
          end;
        end;
        else do;
          if enlYrsON2 > 0 then do sted = 1 to 5;
            enlYrs = enli(sted);
            if dg0 > 0 then do grad = 0 to 1;
              dg = grad;
              reflist = "enlYrs dg";
              OUTPUT; /* enlYrs, dg signficant */
            end;
            else do;
              reflist = "enlYrs";
              OUTPUT; /* enlYrs signficant */
            end;
          end;
          else do;
            if dg0 > 0 then do grad = 0 to 1;
              dg = grad;
              reflist = "dg";
              OUTPUT; /* dg signficant */
            end;
            else do;
              reflist = " ";
              OUTPUT; /* no signficant vars */
            end;
          end;
        end;
      end;
    run;
%mend;
/* calculate covariates for all AFSCs */
%DO_OVER(values=&afsclist2, MACRO=survival_functions);
%macro survival_functions_long;
/***************************************************************************/
/* The vars in this macro must be input manually using the output from the previous*/
/* macro.  The class var list should only include the categorical vars that are */
/* relevant for that CF.  The var list behind the first = in the "model" line of */
/* code must also be updated to only have the relevant vars.  One block of */
/* code should be here for every CF of interest.  In this situation, */
/* there were no significant factors for OSI, so that CF is not included    */
/***************************************************************************/
  ods graphics on;
  proc phreg data=offdata.survive (where=(cf="NRO")) plots(overlay)=survival;
    by cf; /* This creates a different survival line for each afsc */
    class gender commission enlYrs dg; /* Categorical vars to use in the analysis */
    model end_cyos*censor_cf(0) = gender commission enlYrs dg/entry=start_cyos;
        /* list all vars after the = */  /* start_cyos is the left-truncated data */
    baseline covariates=offdata.coef_surv_NRO out=surv_NRO survival=_all_;
  run;
  proc phreg data=offdata.survive (where=(cf="LOG")) plots(overlay)=survival;
    by cf; /* This creates a different survival line for each afsc */
    class gender enlYrs dg; /* Categorical variables to use in the analysis */
    model end_cyos*censor_cf(0) = gender enlYrs dg/entry=start_cyos;
        /* list all vars after the = */  /*start_cyos is the left-truncated data */
    baseline covariates=offdata.coef_surv_LOG out=surv_LOG survival=_all_;
  run;
  proc phreg data=offdata.survive (where=(cf="SPT")) plots(overlay)=survival;
    by cf; /* This creates a different survival line for each afsc */
    class gender dg; /* Categorical variables to use in the analysis */
    model end_cyos*censor_cf(0) = gender dg/entry=start_cyos;
        /* list all vars after the = */  /* start_cyos is the left-truncated data */
    baseline covariates=offdata.coef_surv_SPT out=surv_SPT survival=_all_;
  run;
  proc phreg data=offdata.survive (where=(cf="ACQ")) plots(overlay)=survival;
    by cf; /* This creates a different survival line for each afsc */
```

```
      class commission enlYrs; /* Categorical variables to use in the analysis */
      model end_cyos*censor_cf(0) = commission enlYrs/entry=start_cyos;
         /* list all vars after the = */  /* start_cyos is the left-truncated data */
      baseline covariates=offdata.coef_surv_ACQ out=surv_ACQ survival=_all_;
    run;
/* This will calculate the survival function for the entire NRL. Although OSI's */
/* survival equation is not calculated (or -able), those data are included in NRL */
  proc phreg data=offdata.survive plots(overlay)=survival;
     class gender commission enlYrs dg; /* Categorical vars to use in the analysis */
     model end_cyos*censor_cf(0) = gender commission enlYrs dg/entry=start_cyos;
         /* list all vars after the = */  /* start_cyos is the left-truncated data */
     baseline covariates=offdata.coef_surv_NRL out=surv_NRL survival=_all_;
  run;
  ods graphics off;
/* Create a value for cf in NRL*/
  data surv_NRL;
    set surv_NRL;
    cf="NRL";
  run;
%mend;
%survival_functions_long;
%macro export_surv; /* Export all of the survival functions into one excel file */
  proc export data=surv_NRO outfile=survive DBMS=EXCELCS REPLACE; run;
  proc export data=surv_LOG outfile=survive DBMS=EXCELCS REPLACE; run;
  proc export data=surv_SPT  outfile=survive DBMS=EXCELCS REPLACE; run;
  proc export data=surv_ACQ outfile=survive DBMS=EXCELCS REPLACE; run;
  proc export data=surv_NRL  outfile=survive DBMS=EXCELCS REPLACE; run;
%mend;
%export_surv;
```

Bibliography

1. R. Danzig, "Driving in the dark: Ten propositions about prediction and national security (washington, dc: Center for new american security, october 2011)," *Joint Force Quarterly*, vol. 65, pp. 32–37.

2. J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-event system simulation*. Prentice Hall, Upper Saddle River, NJ, 2010.

3. "Discrete event simulation software." `https://www.arenasimulation.com/`. Last accessed on January 6, 2015.

4. "Simio simulation software." `http://www.simio.com/index.php`. Last accessed on January 6, 2015.

5. E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *PNAS*, vol. 99, pp. 7280–7287, 2002.

6. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis, 5th edition*. John Wiley & Sons, Inc., Hoboken, NJ, 2012.

7. R. G. Miller Jr, *Survival analysis*, vol. 66. John Wiley & Sons, 2011.

8. J. D. Sterman, "System dynamics modeling," *California management review*, vol. 43, no. 4, pp. 8–25, 2001.

9. M. van Diepen and P. H. Franses, "Evaluating chi-squared automatic interaction detection," *Information Systems*, vol. 31, no. 8, pp. 814–831, 2006.

10. R. Hill, J. Miller, and G. McIntyre, "Applications of discrete event simulation modeling to military problems," in *Proceedings of the 2001 Winter Simulation Conference*, pp. 780–788, IEEE, Piscatawny, NJ, 2001.

11. M. P. Gaupp, "Pilot Inventory Complex Adaptive System (PICAS): An Artificial Life Approach to Managing Pilot Retention," Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA, 1999.

12. M. Schneider and M. Somers, "Organizations as complex adaptive systems: Implications of complexity theory for leadership research," *The Leadership Quarterly*, vol. 17, pp. 351–365, 2006.

13. A. Zegers and S. Isbrandt, "The arena career modeling environment: A new workforce modeling tool for the canadian forces," in *Proceedings of the 2010 Summer Computer Simulation Conference*, (San Diego, CA, USA), pp. 94–101, Society for Computer Simulation International, 2010.

14. J. T. Hall, "Forecasting marine corps enlisted attrition through parametric modeling," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 2009.

15. J. Collofello, D. Houston, I. Rus, A. Chauhan, D. Sycamore, and D. Smith-Daniels, "A system dynamics software process simulator for staffing policies decision support," in *System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on*, vol. 6, pp. 103–111, Jan 1998.

16. S. R. Parker and J. A. Marriott, "Personnel forecasting strategic workforce planning: A proposed simulation cost modeling methodology," in *Proceedings of the 1999 Winter Simulation Conference*, (IEEE, Piscatawny, NJ), pp. 1410–1414, 1999.

17. M. F. Dabkowski, S. H. Huddleston, P. Kucik, and D. Lyle, "Shaping senior leader officer talent: How personnel management decisions and attrition impact the flow of army officer talent throughout the officer career model," in *Proceedings of the*

2010 Winter Simulation Conference, pp. 1407–1418, IEEE, Piscatawny, NJ, USA, 2010.

18. T. Demirel, "A Statistical Analysis of Officer Retention in the U.S. Military," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 2012.

19. T. A. Perry, "An analysis of Primary Military Occupational Specialties on Retention and Promotion of Mid-Grade Officers in the U.S. Marine Corps," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 2006.

20. E. L. Conzen, "An analysis of the impact of fully funded graduate education on the retention of Naval Officers," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 1999.

21. J. H. Hall, "An analysis of army dentists using logistic regression: A discrete-time logit model for predicting retention," tech. rep., Army Dental Command, Fort Sam Houston, TX, 2009.

22. C. A. Castro and A. H. Huffman, "Predicting retention rates of us soldiers stationed in europe," tech. rep., ARMY MEDICAL RESEARCH UNIT-EUROPE, 2002.

23. G. D. Gjurich, "A Predictive model of Surface Warfare Officer Retention: Factors Affecting Turnover," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 1999.

24. M. A. Zinner, "U.S. Marine Corps Company-Grade Officer Retention," Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 1997.

25. "Classifying military personnel (officer and enlisted)," *Air Force Instruction 36-2101*, 2013.

26. M. Gibb, "A1PF sustainment model optimization." personal communication, Dec. 2014.

27. "SAS annotated output: Proc logistic." `http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm`. Last accessed on February 17, 2015.

28. "SAS seminar: Introduction to survival analysis in SAS." `http://www.ats.ucla.edu/stat/sas/seminars/sas_survival`. Last accessed on February 17, 2015.

29. S. I. Inc., *SAS/STAT 13.2(R) User's Guide.* SAS Institute Inc., Cary, NC, 2014.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
26–03–2015	Master's Thesis	Oct 2013 — Mar 2015

4. TITLE AND SUBTITLE

Non-Rated Air Force Line Officer Attrition
Rates Using Survival Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Schofield, Jill A., Major, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENS)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-15-M-128

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Dr. Gerald Diaz
HAF/A1PF
1550 W. Perimeter Rd., Rm 4710
Joint Base Andrews NAF Washington, MD 20762-5000
Email: gerald.diaz.civ@mail.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

HAF/A1PF

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A.
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

The Air Force structures its workforce around rank structure and work specialty codes (Air Force Specialty Codes (AFSCs)). The challenge is to grow and manage personnel to fill a variety of skill sets at a variety of ranks over a 20-30 year planning horizon. To ensure that the missions are accomplished while adhering to congressionally-mandated force allocations, the Air Force is continually attempting to "right size" its force by maintaining the correct balance of personnel in each career field. The Air Force conducts its force structure management responsibility by comparing historical attrition rates to current manpower requirements for each AFSC to determine the "optimal" number of officers needed in each accession yeargroup over a 30-year career. Personnel analysts aggregate the individual yeargroup numbers for each AFSC and call this a "sustainment line." In this study, logistic regression was used to determine which factors are significant to predicting non-rated Air Force line officer retention. The variables considered were commissioning yeargroup, gender, source of commission, number of years served as enlisted, career field grouping, and distinguished graduate at commissioning source and all six were significant. All of these factors are included in the survival analysis, which yielded a total of 99 unique survival functions to characterize officer attrition behavior. Each of the survival functions provides a more specific representation of historic behavior that can be used to predict and/or shape future behavior. To best present the data to decision-makers, the unique survival functions must be aggregated after being weighted according to the respective percentage of the populations they represent.

15. SUBJECT TERMS

personnel analysis, survival analysis, sustainment, statistical modeling, logistic regression

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Raymond R. Hill (ENS)
U	U	U	UU	85	**19b. TELEPHONE NUMBER** *(include area code)* (937) 255-3636 x7469 Raymond.Hill@afit.edu

Standard Form 298 (Rev. 8–98)
Prescribed by ANSI Std. Z39.18