3-2002

Identifying Enlisted Stay and Leave Population Characteristics with Discriminant Analysis

Zabrina Y. Hoggard

IDENTIFYING ENLISTED STAY AND LEAVE POPULATION CHARACTERISTICS WITH DISCRIMINANT ANALYSIS

THESIS

Zabrina Y. Hoggard, 1st Lieutenant, USAF

AFIT/GOR/ENS/02-08

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U. S. Government.

AFIT/GOR/ENS/02-08

IDENTIFYING ENLISTED STAY AND LEAVE POPULATION CHARACTERISTICS WITH DISCRIMNANT 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

Zabrina Y. Hoggard, BS

1st Lieutenant, USAF

March 2002

AFIT/GOR/ENS/02-08

IDENTIFYING ENLISTED STAY AND LEAVE POPULATION CHARACTERISTICS WITH DISCRIMNANT ANALYSIS

Zabrina Y. Hoggard, BS
1st Lieutenant, USAF

Approved:

/s/	01 March 2002
John O. Miller, Lt Col (USAF) (Advisor)	date
/s/	01 March 2002
Paul W. McAree, Maj (USAF) (Reader)	date

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Lt. Col. John Miller, and my reader, Maj Paul McAree, for their guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would, also, like to thank my sponsor, Maj Steven Forsythe, from the Air Force Personnel Operations Agency for both the support and latitude provided to me in this endeavor.

I am also indebted to my friends and family for their support during this effort. Special thanks goes to my peers at AFIT for advice and for being there when I needed someone to listen.

Zabrina Y. Hoggard

Table of Contents

Page

List of Figures

Page

List of Tables

AFIT/GOR/ENS/02-08

Abstract

There exist factors that play a major role in an enlisted Airman's decision to either stay on active duty in the Air Force or separate. The current force structure of the U.S. Air Force and increased loss of enlisted personnel is a major concern as we look at maintaining manpower to meet the needs of the Air Force. The Air Force is reacting to this low retention problem by increasing the bonuses for initial enlistments and reenlistments, home basing, increasing quality of life for Air Force personnel with enlisted dormitory plus-ups, and under AEF personnel have increased predictability of deployment.

This thesis provides a method for identifying the variables that most characterize Stay and Leave populations for enlisted Airmen on active duty in the Air Force. Discriminant Analysis is used to identify population characteristics that categorize the two groups. A methodology is constructed that can discriminate between Airmen that stay on active duty military service and Airmen that leave active duty military service.

IDENTIFYING ENLISTED STAY AND LEAVE POPULATION CHARACTERISTICS WITH DISCRIMINANT ANALYSIS

1. Introduction

Background

Since the inception of the United States Air Force (USAF) in 1947, the number of members on active duty has fluctuated with time as seen in Figure 1-1. This fluctuation

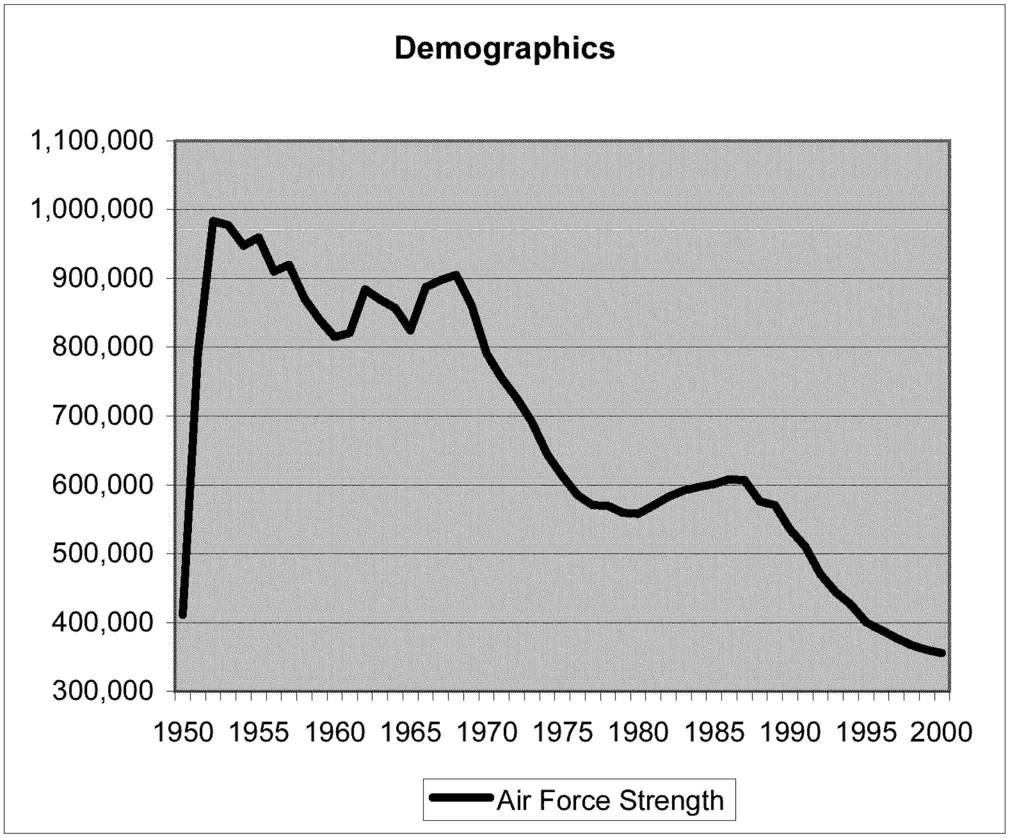


Figure 1-1. Air Force Strength

corresponds to the United States' involvement in military conflicts. The size of the force doubled at the onset of the Korean War, increased again during the Vietnam conflict, and again in Gulf War in the early 1990s. In 1991 Congress voted to trim the United States military by one-fifth over a five-year period. For the Air Force this resulted in a total strength cut from 510,000 to 400,000 from fiscal year 1991 to fiscal year 1995 (Air Force Magazine, February 1991, p.36). This reduction in force dropped total strength to its pre-Korean War level.

From fiscal year 1988 to 1998, the total number of military personnel (end strength) declined by 34 percent, from approximately 2.1 million to 1.4 million. This drawdown occurred in response to the significant changes brought about by the break-up of the Soviet Union and the end of the Cold War. During the initial stages of the drawdown, the Department of Defense (DoD) achieved reductions in end strength largely by limiting accessions (the number of people entering the services). After the Persian Gulf War of 1991, DoD accelerated the drawdown by continuing to limit accessions and instituting voluntary and involuntary separation programs. These programs, which were targeted at service members in different career states and occupations, included authority for early retirements, bonuses for separating from the service or transferring from active duty to the reserves, mandated retirements for people in certain areas who had more than twenty years of service, limitations on reenlistments in areas with personnel surpluses, waivers of service obligations, and reductions in force, (RIFs). To minimize the impact of these programs on the existing force, the Air Force decreased accessions dramatically. The goal of this strategy was to prevent forcing out experienced Airmen with untrained and inexperienced recruits.

Throughout this time of drawdown, when the services were trying to reduce personnel levels, retention was not a primary concern within DoD. However, DoD and Congress have since long recognized that some service members, particularly those in certain technical areas, can be difficult to retain. According to DoD officials, for example, pilots, nuclear engineers and technicians, and medical specialists are all occupations that have experienced retention problems (GAO, 2000).

> Concern with retention increased considerably in 1998, when the services began reporting problems with the readiness and quality of their forces. In September 1998, the Joint Chiefs of Staff testified before the Senate Armed Services Committee that retention had become a top concern, with rates declining both among specific critical personnel such as pilots and naval surface warfare officers and at more aggregate levels, such as among second-term enlisted personnel. This latter finding was a major concern for the services because it implied systematic losses of mid-level, noncommissioned officers. In response to these concerns, Congress passed legislation in 1999 to increase military pay and retirement benefits for service members. The legislation increased base pay for all military personnel, targeted additional pay increases at certain service grades, and repealed legislation providing lower retirement benefits for some personnel. It also required DoD to submit an annual report to Congress on the effects these improvements in compensation and benefits have on recruitment and retention. (GAO, 2000)

For the Air Force, the reduction in the accession levels was predicted to have a dark side: "The Air Force would have enough to meet its needs at the time but barely enough to give it an adequate retention pool when these first-termers become career-eligible Airmen" (January 1992 Air Force Magazine). In the years following the drawdown, this prediction is shown to be true. Retention in the Air Force *is* a major problem and enlisted separation is a major source of concern.

The number of enlisted personnel on active duty in the Air Force shows a significant downward trend in recent years. Demographic data on the total enlisted force numbers lists 462,800 Airmen on active duty in 1989 while in 2000 the number of active duty Airmen on active duty was 282,358; a 38% reduction in just an eleven-year period. Although part of this reduction was planned, the fact that the retention rate continues to decrease is a cause for major concern for the United States Air Force. The Air Force Personnel Operations Agency Enlisted Personnel Analysis Branch (AFPOA/DPYE) is tasked with conducting analysis on retention and reenlistment.

AFPOA/DPYE develops, maintains, and operates a variety of computer models for the analysis of enlisted promotion, retention, accession, compensation, and separations policy alternatives. A suite of human resource models jointly developed by RAND Corporation and Air Force analysts make up the Enlisted Force Management System (EFMS). These models project the aggregated enlisted force, predict impacts of bonus offers, forecast loss rates, forecast reenlistment rates, and determine retention goals. For this research effort we utilize the Loss Model data.

The Loss Model is a regression model. A linear regression equation is estimated to determine propensity to leave the service. (A response variable is 1 if an Airman is lost during the year at risk (YAR), 0 if the Airman is retained at the end of the YAR.) The resulting loss rates are broken down by Air Force Specialty Code, (AFSC), years of service, (YOS), years to expiration term of service, (YETS), grade, and category of enlistment. The loss rates are based on long-term retention trends developed from a 25-year relational database utilizing 24 different linear regression models as predictions for future fiscal years.

Problem Statement

This thesis will concentrate on the analysis of demographic variables and how they affect the separation decision.

Scope

This study focuses explicitly on the data extracted from a SAS data set that is prepared annually to support the EFMS by the Air Force Personnel Center, AFPC. The file used in this study contains longitudinal data on every Airman who was on regular active duty in the Air Force since June 30, 1979. For the purposes of this thesis effort, we will be looking at the most recent observations of enlisted Airmen currently on active duty.

Contribution

The purpose of this thesis is to identify stay and leave populations of enlisted personnel in the Air Force. Through discriminant analysis we will determine if a relationship between demographic variables and a stay/leave decision exists.

Outline

The literature review introduces previous research that motivates the direction of this research by providing information on previous research on retention, explains the data that was available to accomplish the research, and summarizes current tools used in the field of discriminant analysis by providing an overview of current discriminant analysis theory.

Application of discriminant analysis theory is presented in the methodology chapter. The methodology chapter covers how groups were formed with EFMS data, variable selection, and how discriminant analysis theory was applied in this thesis. The fourth chapter presents the results of applying this methodology. In the final chapter conclusions and recommendations will be presented as well as suggestions for follow-on research.

2. Literature Review

The literature review gives detailed information on subjects pertinent to this research effort. The first section discusses previous research in the area of retention and how it relates to our effort. The second section briefly introduces the data that is available for this study. The third section is a detailed explanation of the statistical formulas that are used to perform discriminant analysis.

Previous Research On Retention

There are numerous studies on enlisted retention in the Air Force with some of the studies concentrating on particular subsets of enlisted personnel. In a study on Air Force enlisted aircraft maintenance personnel retention rates, Peter Lommen (1999) found a strong relationship exists between the fluctuations of the economy and the retention of these personnel. Lommen investigated the effects of certain national economic conditions and indicators on enlisted aircraft maintenance personnel retention rates. Specifically, the civilian unemployment rate and the index of eleven leading economic indicators are shown to be excellent predictors of the retention for both first and second term Airmen (Lommen, 1999). Lommen's conclusions parallel with other studies in related fields of interest; for example, the General Accounting Office (GAO) studied the relationship between civilian job opportunities and military retention.

The GAO analyzed the differences between military and civilian compensation in a small sample of occupations. The study showed nearly all of the military occupations had an average pay below their civilian counterparts. Nearly three-quarters of these civilian occupations had their pay fall above the average pay level (GAO, 1986). In a 1992 study by Gill and Haurin, the relationship between a spouse's earnings and the military member's decision to separate is studied. The spouse's income was determined to be lower than that of civilian counterparts due to the mobility associated with the military member. Gill and Haurin concluded this income reduction had a negative effect on the retainability of the military member. This conflict with the spouse's career indicates the influence of economic factors in the stay/leave decision (Gill *et al*, 1992).

Mark Basalla (1996) built a multivariate linear regression model for Air Force personnel management officials that predicts officer retention rates for rated and non-rated line officers. Basalla's research looked at officers split out according to AFSC. He was able to show that in most cases over half of an officer's decision to stay or leave active duty can be explained by economics (Basalla, 1996). However, there have been studies that have looked at non-pecuniary factors when studying retention.

In a 1996 study, Michael Nakada and James Boyle analyzed the effect of the Nuclear Officer Incentive Program (NOIP) on nuclear officer retention for both surface and submarine officers beyond their minimum service requirement. The NOIP provides special pay as a specific retention benefit for nuclear trained officers. Submarine officers can receive $19,000 a year if they obligate for a period no less than 3 years on active duty past their minimum service requirement; this is called COPAY. If the officer chooses no active duty service obligation then the officer will still receive $12,500 for every year in excess of the minimum service requirement; this is called AIB. Once the minimum service requirement is met the officer may choose to: (1) stay in the Navy under contractual obligation (COPAY), (2) stay in the Navy under no contractual obligation

(AIB), or (3) leave the Navy. This study analyzed submarine officers commissioned between fiscal year 1974 and 1989 and showed that increases in NOIP positively influenced retention behavior of the submarine officer at the end of their minimum service requirement (MSR). More pertinent to this research effort, however, it further identified significant differences in retention behavior among different demographic groups (Nakada *et al*, 1996).

For one of the year groups analyzed in the Nakada & Boyle study most of the officers were single at the end of their MSR, and they had the lowest retention rate. If they stayed beyond MSR, single officers were less likely to extend with obligation; while if an officer was married at MSR and/or had more than one dependent at MSR, he was more likely to extend with obligation. The Nakada & Boyle study found that not only does marital status play a role in retention behavior but also as the number of dependents increased, the likelihood the officer stayed on active duty increased, and if he stayed, the likelihood that he was obligated also increased (Nakada *et al*, 1996).

Additional studies have also identified differing retention behaviors among married and single officers. Stephen L. Mehay analyzed performance differences such as promotion, retention and fitness report evaluations between majority and minority Navy and Marine Corp officers. This study used marital status as well as the number of dependents as demographic control variables. Although the study does not account for potential marital status and dependency status changes between the O-3 promotion board and the O-4 promotion board, Mehay was able to find that married officers or those with dependents were more likely to remain in the Navy until the O-4 promotion board (Mehay, 1995).

In response to a significant drop in enlisted second-term retention between fiscal years 1992 and 1994, the Center for Naval Analyses was tasked to develop a predictive model that would relate Navy policy and personal characteristics of enlisted personnel to second-term retention. The analysts utilized various non-pecuniary factors such as civilian unemployment rate, marital status, dependency status, race, and gender in regression models to estimate the impact of each on enlisted retention. The study found that marital status was indeed a significant predictor of enlisted retention decisions. Married sailors were twenty-eight percent less likely to leave after their first enlistment than their single counterparts. Furthermore, regardless of marital status, the probability of leaving after a first-term enlistment decreased with the number of children. In this study an explanation for this retention phenomenon is suggested. The study asserted that married personnel or those with additional dependents might be less likely to undertake disruptive career changes (Moore, 1996).

Retention is not only a problem for the military it is also a problem faced by the private sector. Human Resource managers of businesses and corporations have been battling the issue of retention for quite some time now. Historically, the strategy for retaining key employees was one of raises and promotions. However, as times have changed and employment practices evolved, so have strategies for retaining key personnel. Today other strategies are proving more effective than just monetary reimbursements. As reported by Pam Withers in the July 2001 issue of Workforce, if a company can identify what its employees value then they can build a retention strategy around these values. When answering the question, “Why bone up on employee values?” Withers writes:

> Because if a heated-up economy in the past decade helped thrust so-called knowledge workers into the driver's seat, a slowed-down economy isn't about to eject them. Four other factors have strapped them securely into their position of potency: the global market, worker empowerment, changing demographics, and a determination to make career sacrifices for a better work/life balance. (Withers, 2001)

This rise in global competition and wages has forced human resource directors to turn to a "more organic crop of incentives" most commonly known as soft benefits. Soft benefits, such as increased vacation days, flexible work hours, corporate gym, and in-house childcare, have proven to be more effective and less expensive. The workers of today resist tying up self-identity with their work identity and are looking for ways to balance work and leisure, family and community. Employer retention programs must cater to this shift in values in order to keep their businesses afloat (Withers, 2001).

Another strategy leading the way in retention is to start by attempting to retain the employee even before their first day on the job. This process could start as early as the job orientation. In the past, orientation was seen as a nice thing to do to get the new employee acquainted with their new job. However, today orientation is seen as a critical part of a company's success. These sentiments are expressed in the November 2000 issue of Workforce. Orientation exposes new employees to the company's philosophy and values. This helps them feel like they are a part of the organization they work for and increases their sense of belonging and their commitment from the very beginning. The orientation guidelines should promote such core values as teamwork, communication, creativity, diversity, learning, trust, and quality. If these values and ideals are instilled into the employees daily work life, thereby fostering the desired culture throughout the organization, it is hoped to give them a sense of community and commitment thus increasing retention. CDG and Associates Inc., an organization of consultants who install HRIS systems, utilizes intensive new-hire orientations as a retention tool. A new-hire spends anywhere from one to three weeks in orientation and it seems to be paying off. CDG has a retention rate of more than ninety-three percent. This high retention rate is attributed to the intensive orientation and the firm's nurturing environment by president and founder Cynthia Driskill (Hutchins, 2000).

The two aforementioned private sector retention techniques are just a couple of the new and innovative retention strategies being employed in businesses today. The first suggests finding out what the employees want and catering to their needs in order to keep them happy in hopes of retaining them. The orientation strategy focuses on telling the employee up front what the company values so that they will know what will be expected of them and allow them the opportunity to assess whether the job is right for them. These examples are just the tip of the iceberg when discussing retention strategies.

In a special report on employee retention in the July 2000 issue of HRfocus, retention strategies can be grouped into four categories based on company focus: the basics, pay-related, employee-friendly, and organizational cultural emphasis. The basics category focuses on providing a "family" type professional, informal atmosphere. Employees are allowed to dress informally, help each other as needed, and set their own overtime hours. Pay-related strategies supply employees with pay commensurate to their education level, experience, and job performance. Employee-friendly refers to addressing employee concerns at all levels. Employees want to be listened to, challenged, and recognized for a job well done. This category addresses these issues. In can be expected that no one strategy will work for all, however, the HRfocus survey suggests that the best way to retain employees is to employ a mixed approach in their retention programs. Mixed approaches provide a benefit to both the employer and the employee (HRfocus, 2000).

Many studies have been completed to determine variables affecting job satisfaction and retention. Over the years, studies have also sought to link various variables with employee turnover and have catered retention practices around them. Some of the early studies were restricted to a small number of variables because of the limited availability of data.

Later studies began to identify additional variables, which may affect employee turnover. While the variables affecting turnover were not as important during the drawdown years, these variables have increased in importance as the Air Force seeks to maintain a steady force which can be ready for world wide deployment. This study will continue with analysis of demographic variables suggested to contribute to an Airman's decision to separate.

Available Data

The current data set of active duty Airmen includes 195 variables. Each observation represents an Airman by social security number. For the purpose of this thesis effort only demographic variables will be used in the model.

Discriminant Analysis Overview (Williams, 2001)

We cover three areas of theory in our overview of discriminant analysis: underlying assumptions, classification functions, and assessing results. The underlying assumptions of discriminant analysis require that the groups have a multivariate normal distribution and have equal covariance matrices. There are several different methods of classifying data when performing discriminant analysis. Two methods, Fisher's method and the quadratic discriminant method, are employed in this thesis. Once a classification function is generated, tools are required for assessing classification accuracy and the independent variables that make up the function. We will introduce tools commonly used for this purpose and show how to interpret their results.

Performing discriminant analysis can be summed up in a few simple steps:

1. Check for multivariate normality.
2. Test to see if covariance matrices (Σi for $i = 1,2$) are equal. A common method, Fisher's two-group discriminant analysis, makes this assumption. If this requirement is not met, there are other methods that can be used.
3. Choose a method and compute the discriminant function to generate discriminant scores.
4. Validate the chosen method.

Assessing Multivariate Normality

The first step in assessing multivariate normality is to assess the univariate, or marginal, normality of each independent variable (Andrews, 1974). This can be accomplished with several tools including likelihood tests and normal probability plots. A complete discussion of these techniques can be found in Neter (1996). Although marginal normality does not imply multivariate normality, the presence of most types of deviation from normality will be revealed in the univariate analysis (Andrews, 1974).

If problems are found in the marginal normality, the best solution is to apply a transformation that addresses the specific problem. The most common transformations are exponential and logarithmic transformations for increasing or decreasing variance (Neter *et al*, 1996).

Testing for Unequal Covariance Matrices

The covariance matrix for a multivariate data set is analogous to the variance statistic for a single variable (Giri, 1996). The covariance matrix, labeled Σ, is rarely known for real world data. The sample variance matrix, shown in Equation (1), is used to approximate Σ.

$$S = \frac{1}{n-1} \cdot X_d^T \cdot X_d, \text{ where } X_d = X_i - \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot \overline{X_i}^T \qquad (1)$$

The Bartlett and Box test for unequal covariance matrices is used in this research to determine whether a pooled covariance matrix can be used or not. Giri (1996) and Bauer (2000) show formulations of this test for testing the equality of multiple covariance matrices. The following is a simplified formulation for use in a two-group problem applicable to this research effort.

Let X_1 and X_2 be matrices of independent variables corresponding to two *a priori* defined groups. Now define A_i to be the crossproduct of the mean corrected data for i = 1, 2.

$$A_i = \left[X_i - \begin{bmatrix} 1 \\ \cdots \\ 1 \end{bmatrix} \cdot \overline{X_i}^T \right] \cdot \left[X_i - \begin{bmatrix} 1 \\ \cdots \\ 1 \end{bmatrix} \cdot \overline{X_i}^T \right]^T \quad i = 1,2 \qquad (2)$$

Also Define $A = A_1 + A_2$

p = # of variables
n_i = sample size in group *i*
$N = n_1 + n_2$
$n = N - 2$

The test for equal covariance matrices is a standard hypothesis test. A test statistic is found and is compared to a χ^2 distribution where α is the chance of type I error with $1/2 \cdot p \cdot (p-1)$ degrees of freedom.

Null Hypothesis: $\Sigma_1 = \Sigma_2$

Rejection Region: test statistic > $\chi^2_{1-\alpha, \frac{1}{2} \cdot p \cdot (p-1)}$

The test statistic is:

$$\omega 2 = -2 \cdot \rho \cdot \ln(W) \tag{3}$$

where $$\rho = 1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n} \right) \cdot \frac{(2 \cdot p^2 + 3 \cdot p - 1)}{6 \cdot (p+1) \cdot (q-1)}$$

$$W = e^V \cdot \left[(\frac{n}{n_1 - 1})^{\frac{(p \cdot (n_1 - 1))}{2}} \cdot (\frac{n}{n_2 - 1})^{\frac{(p \cdot (n_2 - 1)}{2}} \right]$$

$$V = \left[\frac{(n_1 - 1)}{2} \cdot \ln(|A_1|) + \frac{(n_2 - 1)}{2} \cdot \ln(|A_2|) - \frac{n}{2} \cdot \ln(|A|) \right].$$

These are complicated formulas that can be difficult to compute. Guri (1996) shows a much simpler approximation for the test statistic derived by Box shown in

Equation (4). Here ρ is the term defined in Equation (3) and S_i is the covariance matrix of the ith group and S_p is the pooled covariance matrix defined in Equation 5:

$$\omega 2 = \rho \cdot \left[n \cdot \ln\left(\left|S_p\right|\right) - (n_1 - 1) \cdot \ln\left(\left|S_1\right|\right) - (n_2 - 1) \cdot \ln\left(\left|S_2\right|\right) \right] \qquad (4)$$

$$S_p = \frac{1}{n_1 + n_2 - 2} \cdot \left[(n_1 - 1) \cdot S_1 + (n_2 - 1) \cdot S_2 \right] \qquad (5)$$

Classification Methods

One goal of a classification method is to assign a scalar score to each object in the data set. The score determines class membership. Two classification methods are covered in this section, Fisher's linear approach and the quadratic discriminant function. Fisher's approach has a body of supporting literature and is simple to compute. The quadratic discriminant function is a more robust classifier in that it does not require equal covariance matrices.

Fisher's approach does not make any distribution assumptions for the variables. This method finds a linear combination of the object's attributes with the goal of maximizing the distance between the means and minimizing the variance (Dillon, 1984). Fisher proved that Equation (6) forms such a linear combination.

$$\overline{b} = \mathbf{S}_p^{-1} \cdot (\overline{\mathbf{X}_1} - \overline{\mathbf{X}_2}) \qquad (6)$$

Where S_p is the sample pooled covariance matrix.

This equation produces a vector of weights. Scores for each object are calculated by multiplying each object attribute by the appropriate weight. An entire group's scores can be calculated with Equation (7),

$$score_i = X_i \cdot \left(\overline{b} \right), \text{ for group i} \tag{7}$$

After scores have been calculated, a classification rule is imposed. For Fisher's method, the simplest rule is using the midpoint of the scores as a dividing point. First determine the mean score for both groups and then determine the overall mean of the scores. Call the group with mean score smaller than the overall mean group A and the group with score larger than the overall mean group B. New objects with scores smaller than the overall mean are classified as group A and objects with scores larger than the overall mean are classified as group B. Because this classification method will not achieve 100% accuracy we must consider the misclassification error percentage. The misclassification error percentage as calculated by applying the discriminant function to the data that was used to formulate the weights is called the apparent error-rate. The following is a step-by-step guide to determining the apparent error rate for Fisher's method. This method only works for groups with equal number of objects.

1. Determine the means of the scores for groups 1 and 2.
2. Find the mean of the combined scores using Equation (8).

$$\text{Midpoint } = 0.5 \cdot (\overline{X_1} - \overline{X_2})^T \cdot S_p \cdot (\overline{X_1} + \overline{X_2}) \tag{8}$$

3. Determine scores for X_1 and X_2 using Equation (7).

4. Let the number of correctly classified objects for X_1 be c_1 and likewise let the number of correctly classified objects for X_2 be c_2.

5. Calculate the apparent error-rate (APER) as the ratio of misclassified objects to the total number of objects.

$$\text{APER} = \frac{(n_1 - c_1) + (n_2 - c_2)}{n_1 + n_2}. \qquad (9)$$

One advantage of using Fisher's method is the existence of a statistical test to assess the Mahalanobis distance between the means of the groups when the covariance matrices of the two groups are equal (Giri, 1996). This allows for a quick test of data for the likelihood of successful discriminant analysis. The following is a description of Hotelling's T^2 statistical test:

Null Hypothesis: $\mu_1 = \mu_2$

Rejection Region: Test statistic > $F_{(1-\alpha, n_1+n_2-p-1)}$

$$\text{Test Statistic:} \quad \frac{n_1 + n_2 - p - 1}{p \cdot (n_1 + n_2 - 2)} \cdot T^2 \qquad (10)$$

$$\text{where} \quad T^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} \cdot \left(\overline{X_1} - \overline{X_2}\right)^T \cdot S_p^{-1} \cdot \left(\overline{X_1} - \overline{X_2}\right).$$

Quadratic discriminant scores require the underlying assumption of multivariate normality and allow for groups with unequal covariance matrices (Dillon, 1984).

Another attribute of the quadratic discriminant scores approach (D_q scores) is we can separate groups that are not linearly separable (Bauer, 2000). The disadvantage of using the D_q scores approach is the higher level of computational complexity.

The D_q scores approach is an approximation of the natural log of the likelihood estimator. Each object classified receives a score from the likelihood estimator for the first group and a score from the likelihood estimator for the second group. The formula for calculating the quadratic discriminant scores for an object i from matrix X is,

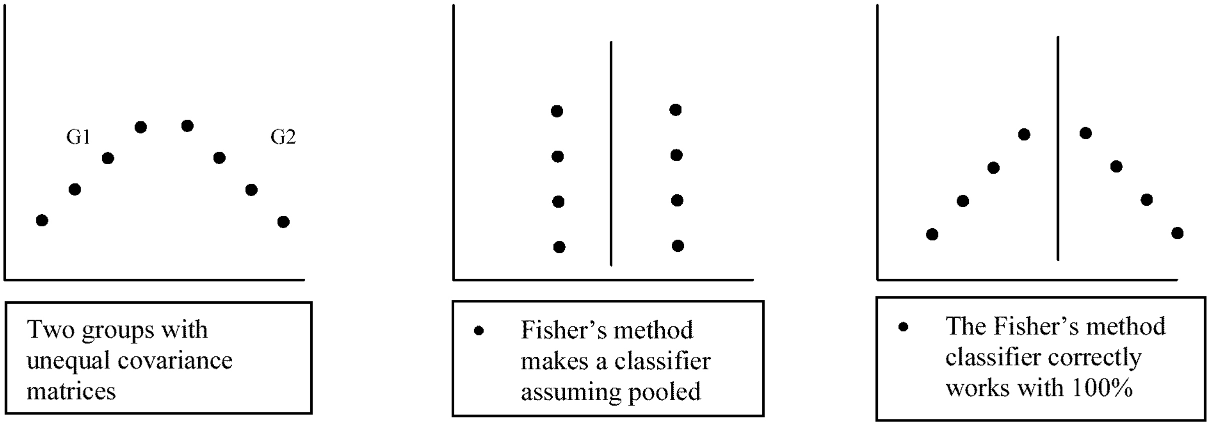
$$\begin{aligned} D_1^q(X_i) &= -\frac{1}{2}\cdot\ln\left|S_1\right| - \frac{1}{2}\cdot(X_i - \overline{X_1})^T \cdot S_1^{-1}\cdot(X_i - \overline{X_1}) + \ln(\frac{n_1}{n_1+n_2}) \\ D_2^q(X_i) &= -\frac{1}{2}\cdot\ln\left|S_2\right| - \frac{1}{2}\cdot(X_i - \overline{X_2})^T * S_2^{-1}\cdot(X_i - \overline{X_2}) + \ln(\frac{n_2}{n_1+n_2}) \end{aligned} \qquad (\ 11\ )$$

After both scores have been calculated the object is classified into the group with the larger D_q score (higher likelihood).

Quadratic discriminant scores are similar to Fisher discriminant scores in that they will not necessarily correctly classify 100% of all objects. An apparent error rate can be calculated with Equation (9).

Deciding on a Classification Method

Deciding on a classification method is not a straightforward decision. If we assume the covariance matrices as equal and do not make any distribution assumptions, then Fisher's method appears to be appropriate. Otherwise we will apply the Dq scores approach. Unfortunately, application may not prove that straightforward. Some groups with unequal covariance matrices can be correctly classified by Fisher's method if the means of the groups are far enough apart. This is shown in Figure 2-1.



Two groups with unequal covariance matrices

- Fisher's method makes a classifier assuming pooled

- The Fisher's method classifier correctly works with 100%

Figure 2-1. The Flexibility of Fisher's Method

We recommend using both methods to classify the data instead of depending solely on the condition of the covariance matrices to guide the decision. We will incorporate into the decision the performance of the discriminant analysis procedure by measuring the apparent error rate.

Variable Contribution

When a discriminant function derives a classification function, it makes use of all the data available. Determining which variables make the greatest contribution to classifying the data is the main focus of this research. The score generated by the discriminant function is an artificial variable that, by design, is the most effective tool for distinguishing between the defined groups. The independent variables that have the highest magnitude of correlation with the discriminant scores, arguably, have the most contribution to classifying the data. Discriminant loadings are the correlations of the independent variables with the scores produced by the chosen discriminant function (Dillon, 1996). Figure 2-2 illustrates how variables can exhibit varying influence on classification.

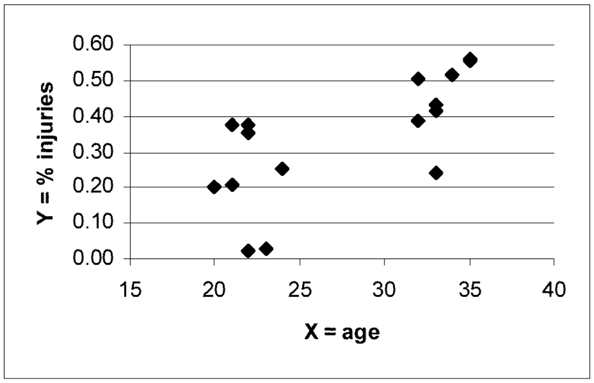


Figure 2-2. Understanding The Need For Loadings

In this example, two variables have been collected by the researcher, customer age, X, and percentage of injuries, Y. Fisher's method produces the classification function $-7.7 \cdot X - 4.0 \cdot Y$. Because the magnitude of the coefficents are similar, a researcher may infer that both variables are important for classification. This is a false assumption caused by the different scales of the two variables. Discriminant loadings give a more accurate assesment because they are a correlation measurement and therefore unitless. The loadings for this problem show a correlation of -0.999 for X and -0.189 for Y with the discriminant scores. Age is clearly the more important variable for classification as reflected in the loadings.

When computing discriminant loadings for groups with equal covariance matrices, as is the case in Fisher's method, Equation (12) will produce a vector of the correlations of each variable with the discriminant function.

Define

$$D_x = \begin{bmatrix} \frac{1}{\sqrt{S_{i,i}}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{S_{i,i}}} \end{bmatrix}$$

$$\bar{b} = S^{-1} \cdot (\overline{X_1} - \overline{X_2})$$

$$D_{bx} = \left(\bar{b}^T \cdot S \cdot \bar{b}\right)^{-\frac{1}{2}}$$

$$loadings = D_{bx} \cdot D_x \cdot S \cdot \bar{b} \qquad (\ 12\ )$$

If the groups do not have equal covariance matrices, loadings can still be computed. Computing a univariate correlation of each variable with the discriminant scores one at a time with Equation (13) provides a loading vector.

$$\bar{r} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \qquad (\ 13\ )$$

Summary

This literature review examined some of the available research today dealing with retention issues. Previous research suggests that as a result of the drawdown, today's military is faced with shortages in key career fields. These shortages affect the ability of the Air Force to effectively conduct its mission. In general, the previous studies show that research into the evaluation of demographic variables and their impact on the separation decision may prove beneficial for future retention policies.

3. Methodology

Introduction

The underlying goal of this research is to use a data mining technique called discriminant analysis to see if trends exist within the available datasets. It is hoped that identification of these trends will allow for an interpretation about the retention behaviors of enlisted personnel. This chapter explains how the SAS datasets made available by AFPOA were analyzed. It covers independent variables, group definition, underlying assumptions, describes calculation of the discriminant loadings, and variable selection.

Independent Variables

The original datasets contained 195 variables as shown in Table 3-1. These variables represented factors about each Airman such as social security number, career field groups, bonus information, and demographic variables. Descriptions of each variable in the entire list are not given due to the fact that most of the variables are solely used for purposes internal to AFPOA. Descriptions were also deemed unnecessary due to the fact that only the demographic variables will be utilized in this research effort.

The independent variables for this research are the demographic variables supplied in the datasets and are highlighted in bold type. The twenty-three demographic variables used in this research are for the most part categorical in nature (19 out of 23) and include information such as marital status, age at time of enlistment, minority status, and level of education. We will be using these variables to assign observations into one of the two groups: Stay or Leave. The computer program SAS will be utilized to perform the discriminant analysis.

Table 3-1. Original Variables

1	**ssan**	40	extlen	79	GRADE7	118	enddate2	157	edats12
2	stoploss	41	HYT	80	GRADE8	119	cnvindt1	158	badswaf12
3	yets	42	EO	81	**AFQTGP1**	120	cnvindt3	159	cjrres12
4	yyos12	43	PC	82	**AFQTGP2**	121	cnvindt2	160	bonus18
5	tnts	44	HIMISC	83	**AFQTGP3**	122	cnvindt4	161	bon180_5
6	yriskb	45	VSIEXP93	84	TERM6	123	previus1	162	bon185_1
7	tbend	46	VSIEXP94	85	M1	124	previus3	163	bon181_p
8	txoets	47	VSIEXP95	86	M2	125	previus2	164	delt1812
9	tprvbm	48	VSILOS93	87	M3	126	previus4	165	up1812
10	pafsc	49	VSILOS94	88	M4	127	numfeed1	166	down1812
11	uafsc	50	VSILOS95	89	M5	128	numfeed3	167	cjrres18
12	TIME	51	HYTPROM	90	M6	129	numfeed2	168	bonus24
13	EXTTIME	52	EMPID	91	M7	130	numfeed4	169	delt2418
14	ETSPRIME	53	CFG	92	M8	131	prevbeg1	170	up2418
15	acc_fy	54	RANDCFG	93	M9	132	prevbeg3	171	down2418
16	**LOST**	55	FIRCFG	94	M10	133	prevbeg2	172	cjrres24
17	EXTEND	56	DODCFG	95	M11	134	prevbeg4	173	rate1812
18	REENL	57	mets	96	**DEPEND1**	135	eligible	174	rate2418
19	OVERSEA	58	reup12	97	**DEPEND2**	136	zone	175	x12rat3
20	**MALE**	59	lagren12	98	one	137	cfgnew	176	x12rat4
21	FEMALE	60	lagssn	99	cfg1	138	yarend	177	lossrat3
22	**BLACK**	61	ext12	100	cfg2	139	trim	178	lossrat4
23	**WHITE**	62	payrate	101	cfg3	140	tim2	179	rerat4
24	**MINORITY**	63	unemply	102	cfg4	141	time12	180	x12rate4
25	AFQTGP	64	CJRREST	103	cfg5	142	pafsc12	181	lossrait4
26	**MARRY**	65	fy12	104	cfg9	143	time18	182	rerat3
27	SINGLE	66	PCTBLK	105	**UNDR18**	144	pafsc18	183	x12rate3
28	**NOHSG**	67	PCTMINOr	106	**AGE18**	145	time24	184	lossrait3
29	**GED**	68	PCTFEM	107	**OVER18**	146	pafsc24	185	cfgr1812
30	**HSGRD**	69	PCTTOE6	108	ruafsc	147	bonus12	186	cfgr2418
31	**SOMCOL**	70	PCTAFQT1	109	ruafscm	148	bonus0	187	cfgext3
32	**COLGRD**	71	PCTAFQT2	110	xuafsc	149	bon120_5	188	cfgext4
33	NOTHN	72	PCTAFQT3	111	xuafscm	150	bon125_1	189	cfgloss3
34	**GRADE**	73	size	112	luafsc	151	bon121_p	190	cfgloss4
35	**TERM**	74	swafdate	113	luafscm	152	bdat	191	x12dep
36	DEPEND	75	GRADE3	114	not3p	153	edat	192	lostdep
37	ACC_MON	76	GRADE4	115	begdate	154	swafs12	193	cutoff
38	Enlage	77	GRADE5	116	pafsc2	155	beg	194	lastyar
39	**yos**	78	GRADE6	117	eday2	156	bdats12	195	yar

Defining *a priori* Groups

An *a* priori group is a collection of two sets that are mutually exclusive and exhaustive. The term "Mutually exclusive and exhaustive" implies that each object that is studied is assigned to one and only one of the groups (Dillon *et al*, 1984). Figure 3-1 is an example of the easiest way to do this: define groups with an indicator variable.

An example significant to our research would be looking at the difference in retention behavior of Airmen based on sex. Each record in the study represents a service member and can be assigned to one and only one group.

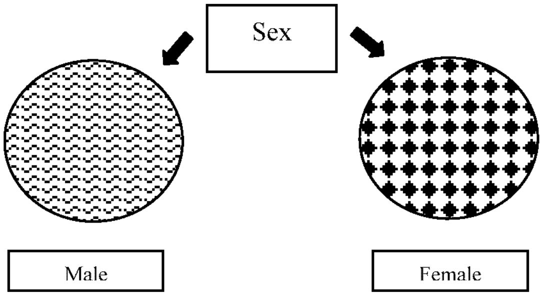


Figure 3-1. Defining *a priori* Groups (Williams, 2001)

Discriminant analysis can be used on datasets that utilize continuous variables as well. Knowing which grouping method to use requires a careful study of the data and a firm grasp of the problem that needs to be solved. Group definition is a flexible process as long as the groups that are formed are mutually exclusive and exhaustive. Nevertheless, the datasets used in this study do not contain any continuous variables and therefore we will use the indicator variable approach. Our research goal deals with stay and leave populations so LOST is our indicator variable. Therefore, for the purpose of this research the two groups are defined as those who stay on active duty (LOST=0) and those who leave active duty (LOST=1).

Underlying Assumptions

It is important to note here that in this research the assumption of equal covariance matrices is not met nor is the assumption of multivariate normality. A large number of studies have utilized discriminant analysis while violating these assumptions; this supports the contention that robustness is not a problem as stated in the text Multivariate Analysis Methods and Applications (Dillon *et al*, 1984). Although the underlying assumptions are not met, robustness of the technique provides information but the results are not statistically rigorous.

Calculating Quadratic Discriminant Score Loadings (Williams, 2001)

In Chapter II it was shown that Equation (12) could be used to calculate loadings for groups with equal covariance matrices. Chapter II also suggested that loadings could be calculated for problems that use quadratic discriminant scores to classify the groups. There are different ways to calculate these loadings. This section demonstrates how the quadratic discriminant (D_q) score loadings were calculated in this thesis.

When using D_q scores to classify the data there are two scores calculated for every object. The first, D_{q1}, is an estimate to the likelihood that the object is from the first group. The second, D_{q2}, is also an estimate, this time for the likelihood that the object is from the second group. The object is assigned to the group with the largest likelihood as seen in Figure 3-2. In Chapter II it was suggested that the loadings could be calculated one variable at a time with the univariate correlation formula. At this point there are two discriminant scores, D_{q1} and D_{q2}. These scores have to be combined in order for the calculation to be performed. Equation (14) is the composite discriminant score, D_{qc}, formed by subtracting D_{q2} from D_{q1}. Calculating the univariate correlation of each variable with the composite quadratic discriminant score formed the quadratic discriminant loadings.

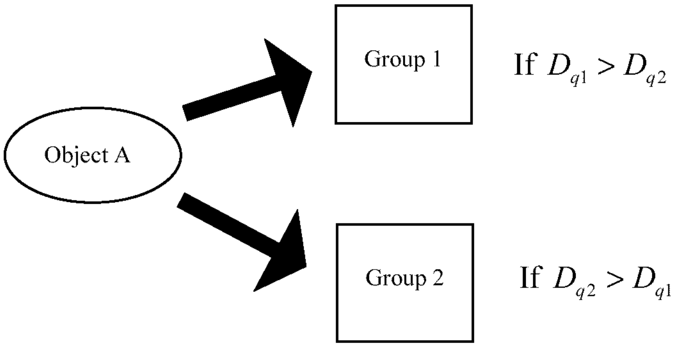


Figure 3-2. Classifying Data With D_q Scores

$$D_{q1} - D_{q2} = D_{qc} \tag{14}$$

Variable Selection

The first step of this process is to determine what combination of the twenty-three variables will be used in the analysis avoiding interaction among variables. It is more difficult to explain variable interaction than single variables. It is desirable to avoid interactions if an accurate classification function can be formed without them. After the data has been classified by SAS we will use the Resubstitution Method of estimating errors of misclassification (Dillon, 1984). The Resubstitution Method produces the *apparent error rate*, APER. The APER is calculated by summing the number of misclassifications from each group and dividing by the sum of both the misclassifications and correct classifications for each group. Equation (15) is an adaptation from Chapter

2 of Equation (9). Table 3-2 is the confusion matrix showing where the values are obtained (Bauer, 2001):

$$\text{APER} = \frac{N_{\bar{0}} + N_{\bar{1}}}{N_0 + N_{\bar{0}} + N_1 + N_{\bar{1}}} \qquad (15)$$

Table 3-2. Confusion Matrix

		PREDICTED MEMBERSHIP	
		Stay	**Leave**
ACTUAL MEMBERSHIP	**Stay**	N_0	$N_{\bar{1}}$
	Leave	$N_{\bar{0}}$	N_1

where N_i = number of group i classified correctly and $N_{\bar{i}}$ = number of group i misclassified for i = 0 (Stay), 1 (Leave).

Once an acceptable error rate is achieved the researcher will be able to analyze the results based on the variables used when obtaining this APER. For this research it is our goal to obtain a minimum APER, as close to zero as possible. A high APER does not give the researcher confidence to recommend acceptance of this method and its use as an accurate classification method based on the variables used.

With the twenty-three variables it is possible to perform discriminant analysis with each variable taken one at a time, two at a time, three at a time, and so on until all twenty-three at once. This process produces a significantly large number of variable combinations. One of the challenges of this research is to determine the variable combinations that return the lowest APER without attempting to perform analysis with every possible variable combination. This begins the “art” of the analysis process.

Variable combination selection continues until the APER reaches a minimum. The smallest set of variables that produce an APER closest to 0 is the set that could potentially be used to classify future data.

Summary

This methodology introduced the variables provided in the datasets, the twenty-three demographic independent variables that will be used in the analysis were identified, and we explained how the groups would be chosen. We also discussed the underlying assumptions and their affect on the analysis. As the method for variable selection was described the apparent error rate was introduced as the measure for assessing the results. In the next chapter we will present the analysis of the variables and display the results.

4. Data Analysis and Results

This chapter describes the analyses conducted on the SAS data set using the discriminant analysis technique. There are three files used in the analysis. The first contains information on first-term Airmen, the second contains information on second-term Airmen, and the third contains information on career-term Airmen. The first few sections of this chapter will explain the process used on all three data sets and then a section on each individual dataset will follow. Other sections will discuss items learned in the process of the research and the supporting analysis.

It is important to note here that the tables presented in this chapter represent output from the data mining process. Data mining is also known as KDD or Knowledge Discovery in Databases. KDD is defined as "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad *et al*, 2001). At this point in the research the goal is to creatively and systematically produce output that will allow for interpretation later. This is not an easy process and is very much an art as well as a science. The approach used during this exploratory process was to first concentrate on only one dataset. Once a course of action was achieved this methodology would be applied to the subsequent datasets. As with any research the learning process is not necessarily chronological. That being said, be advised that along with the aforementioned approach if any new knowledge was gained while analyzing subsequent dataset these discoveries were in the end utilized on all datasets.

Table 4-1. Variable Descriptions

Variable:	Type	Description
AFQTGP1	0-1	AFQT Group I (93<X<99)
AFQTGP2	0-1	AFQT Group II (65<X<92)
AFQTGP3	0-1	AFQT Group IIIA (50<X<64)
AGE18	0-1	Age = 18 at first enlistment
OVER18	0-1	Age > 18 at first enlistment
UNDR18	0-1	Age < 18 at first enlistment
BLACK	0-1	Race = Black
WHITE	0-1	Race = Caucasian
MINORITY	0-1	Race = Neither Black nor Caucasian
COLGRD	0-1	College Graduate
HSGRD	0-1	High School Diploma
NOHSG	0-1	No High School Diploma
GED	0-1	Possesses GED
MALE	0-1	Code 0 for female, 1 for male
MARRY	0-1	1 if married, 0 otherwise
DEPEND1	0-1	One dependent
DEPEND2	0-1	Two or more dependents
GRADE	E-X	Airman's pay grade
LOST	0-1	1 if left active duty, 0 if stayed
SOMCOL	0-1	Some College Training
TERM	Numeric	4 if 4-year term, 6 if 6-year term
SSAN	Numeric	Social Security Number
YOS	Numeric	Number of years of service

Dataset Manipulations

In order to perform the discriminant analysis procedure the datasets were manipulated into a more easily useable state. All variables not being used in the discriminant analysis procedure were removed. This decreased the size of the files considerably and allowed for faster computation. All the records were checked to ensure that there were no duplicate social security numbers. Each observation represents an individual Airman. All records that contained missing data were deleted. In SAS the discriminant analysis procedure automatically omits records with missing data from analysis, however, these records were deleted to decrease the size of the files and increase computation speed. Under the variable AFSC, the data was reduced to only the first three letters.

First-term Airmen Analysis

The initial dataset consisted of 299,418 records and the reduced dataset contains 292,922 records. A discriminant analysis procedure was run in SAS on each variable individually. Table 4-2 shows the results of the discriminant analysis procedure sorted by APER from lowest to highest. We obtained the lowest APER with the variable GRADE. This suggests that GRADE is a variable providing the greatest predictability. The next step was to see if a lower APER could be achieved by increasing the number of variables used in the analysis while including the variable that had originally given a low APER.

Table 4-3 displays results of the discriminant analysis procedure for two variable combinations involving GRADE sorted by APER from lowest to highest. From this we are able to see that we obtained a lower APER when the variable GRADE was paired with each of the variables: BLACK, YOS, TERM, and DEPEND2.

Table 4-4 displays results of the discriminant analysis procedure for three variable combinations involving GRADE and BLACK sorted by APER from lowest to highest. From this we are able to see that no lower APER was obtained when the variables GRADE and BLACK were grouped with each of the remaining relevant variables.

Table 4-2. APER for First-Term Airmen, Variables Taken One at a Time

	n0=# of actual stays	n1=#of actual leaves	
	182854	110068	
VARIABLES	0 missclass	1 missclass	APER
	classified as a Leave but really is a Stay	classified as a Stay but really is a Leave	
GRADE	27554	69944	0.33285
NOHSG	334	109834	0.37610
GED	3668	107815	0.38059
AFQTGP1	9303	103766	0.38600
AGE18	56733	74935	0.44950
OVER18	63806	70768	0.45942
HSGRD	67399	69488	0.46732
AFQTGP2	79058	60341	0.47589
MARRY	99356	41282	0.48012
AFQTGP3	91473	52449	0.49133
WHITE	142298	16414	0.54182
SOMCOL	123835	35465	0.54383
BLACK	151739	11210	0.55629
DEPEND1	140653	23098	0.55903
DEPEND2	153026	12050	0.56355
MALE	150952	17544	0.57522
TERM	162365	7774	0.58083
YOS	164332	8249	0.58917
MINORITY	173413	5204	0.60978
UNDR18	175781	4167	0.61432
COLGRD	178473	2628	0.61826

Table 4-3. APER for First-Term Airmen, Variables Taken Two at a Time

		n0=# of actual stays	n1=#of actual leaves	
		182854	110068	
		0 missclass	1 missclass	
Variables		classified as a Leave but really is a Stay	classified as a Stay but really is a Leave	APER
GRADE	BLACK	22507	74325	0.33057
GRADE	YOS	27127	70023	0.33166
GRADE	TERM	27183	70133	0.33222
GRADE	DEPEND2	27494	69954	0.33268
GRADE	AFQTGP2	27554	69944	0.33285
GRADE	AFQTGP3	27554	69944	0.33285
GRADE	AGE18	27554	69944	0.33285
GRADE	COLGRD	27554	69944	0.33285
GRADE	DEPEND1	27554	69944	0.33285
GRADE	HSGRD	27554	69944	0.33285
GRADE	MALE	27554	69944	0.33285
GRADE	MARRY	27554	69944	0.33285
GRADE	MINORITY	27554	69944	0.33285
GRADE	OVER18	27554	69944	0.33285
GRADE	SOMCOL	27554	69944	0.33285
GRADE	UNDR18	27554	69944	0.33285
GRADE	WHITE	27554	69944	0.33285
GRADE	NOHSG	27698	69884	0.33313
GRADE	GED	29541	69143	0.33690
GRADE	AFQTGP1	36104	65269	0.34608

Table 4-4. APER for First-Term Airmen, Variables Taken Three at a Time

				n0=# of actual stays	n1=#of actual leaves	
				182854	110068	
				0 missclass	1 missclass	
	Variables			classified as a Leave but really is a Stay	classified as a Stay but really is a Leave	APER
1	GRADE	BLACK	AFQTGP2	22507	74325	0.33057
2	GRADE	BLACK	AFQTGP3	22507	74325	0.33057
3	GRADE	BLACK	AGE18	22507	74325	0.33057
4	GRADE	BLACK	DEPEND1	22507	74325	0.33057
5	GRADE	BLACK	HSGRD	22507	74325	0.33057
6	GRADE	BLACK	MALE	22507	74325	0.33057
7	GRADE	BLACK	OVER18	22507	74325	0.33057
8	GRADE	BLACK	SOMCOL	22507	74325	0.33057
9	GRADE	BLACK	UNDR18	22507	74325	0.33057
10	GRADE	BLACK	NOHSG	22656	74260	0.33086
11	GRADE	BLACK	GED	24283	73558	0.33402
12	GRADE	BLACK	COLGRD	25658	72493	0.33508
13	GRADE	BLACK	AFQTGP1	30823	69699	0.34317
14	GRADE	BLACK	MARRY	89721	36137	0.42966
15	GRADE	BLACK	DEPEND2	129727	18463	0.50590
16	GRADE	BLACK	TERM	134631	18109	0.52144
17	GRADE	BLACK	YOS	136459	18228	0.52808

This analysis of First-Term Airmen showed that while certain variables proved lower APERs than others, once further analysis was conducted there was nothing significant that stood out. For example, using the above three tables provided we see that the variable GRADE gives a low APER. Expanding two-dimensionally on this variable proves to be ineffective. The APERs listed in Table 4-3 are all quite similar, with only a difference of 0.1551 from the lowest to highest. Further analysis shows that three-dimensional expansion does not provide significant insight. Next, this same process is applied to the Second-Term dataset. It is hoped that this process will lead to a successful process of identifying patterns in the data.

Second-term Airmen Analysis

Table 4-5. APER for Second-Term Airmen, Variables Taken One at a Time

	n0=# of actual stays	n1=#of actual leaves	
	124053	29677	
VARIABLES	0 missclass	1 missclass	APER
	classified as a Leave but really is a Stay	classified as a Stay but really is a Leave	
NOHSG	118	29637	0.19355
GED	3905	28612	0.21152
AFQTGP1	6381	27808	0.22240
MALE	18315	24346	0.27751
MARRY	31178	20749	0.33778
DEPEND1	30600	21857	0.34123
NUMOFDEP	39925	18247	0.37840
AGE18	39415	20179	0.38765
GRADE	45979	14006	0.39020
AGELEVEL	45853	18657	0.41963
OVER18	45853	18657	0.41963
AFQTGP2	50784	16577	0.43818
SOMCOL	52489	16859	0.45110
TESTSCORELEVEL	57165	14286	0.46478
EDLEVEL	56672	15921	0.47221
AFQTGP3	59860	14286	0.48231
HSGRD	60695	14816	0.49119
DEPEND2	66409	12060	0.51043
YOS	68942	11419	0.52274
TERM	88780	7438	0.62589
RACE	95657	5507	0.65806
WHITE	95657	5507	0.65806
BLACK	101241	4288	0.68646
UNDR18	117615	1522	0.77498
MINORITY	118469	1219	0.77856
COLGRD	119870	938	0.78585

The initial dataset consisted of 158,557 records and the reduced dataset contains 153,730 records. A discriminant analysis procedure was run in SAS on each variable individually. Table 4-5 shows the results of the discriminant analysis procedure sorted by APER from lowest to highest.

Table 4-6. APER for Second-Term Airmen, Variables Taken Two at a Time

		n0=# of actual stays	n1=#of actual leaves	
		124053	29677	
		0 missclass	1 missclass	
Variables		classified as a Leave but really is a Stay	classified as a Stay but really is a Leave	APER
NOHSG	AFQTGP2	118	29637	0.19355
NOHSG	AFQTGP3	118	29637	0.19355
NOHSG	AGE18	118	29637	0.19355
NOHSG	BLACK	118	29637	0.19355
NOHSG	OVER18	118	29637	0.19355
NOHSG	UNDR18	118	29637	0.19355
NOHSG	WHITE	118	29637	0.19355
NOHSG	MINORITY	118	29637	0.19355
NOHSG	DEPEND1	118	29637	0.19355
NOHSG	DEPEND2	118	29637	0.19355
NOHSG	SOMCOL	118	29637	0.19355
NOHSG	TERM	118	29637	0.19355
NOHSG	YOS	118	29637	0.19355
NOHSG	RACE	118	29637	0.19355
NOHSG	EDLEVEL	118	29637	0.19355
NOHSG	AGELEVEL	118	29637	0.19355
NOHSG	TESTSCORELEVEL	6492	27348	0.22013
NOHSG	AFQTGP1	6492	27770	0.22287
NOHSG	MALE	18421	24307	0.27794
NOHSG	NUMOFDEP	25302	21855	0.30675
NOHSG	MARRY	31269	20719	0.33818
NOHSG	GRADE	46072	13984	0.39066

Results similar to those for the First-Term Airmen dataset were reachieved with the Second-Term Airmen dataset. Time was taken to reevaluate the approach to see if there was a way to better represent the data, which would give greater perspective.

Career-term Airmen Analysis

The initial dataset consisted of 180,279 records and the reduced dataset contains 172,249 records. Because of the insight gained while analyzing the First-Term and Second-Term datasets not much preliminary analysis was performed on the Career-Term dataset. Tables describing the analysis on this dataset are included in the sections that follow.

Variable Reduction

The specific request by AFPOA of the research effort was to look at demographic data. This required a severe reduction of the initial 195 variables. The initial reduction left twenty-three demographic variables, Table 4-1. A series of test runs were accomplished with these variables to get preliminary analysis and to provide direction for continued variable selection. These preliminary runs, presented in previous sections, provided a wealth of information that lead to further variable grouping and the consideration of other variables other than the initial twenty-three. The variable AFSC, (Air Force Specialty Code), was added as it may provide greater insight.

Variable Groupings

Associated variables were logically grouped in an effort to gain more insight into possible predictability. The variables BLACK, WHITE, and MINORITY were grouped under one variable entitled RACE. The variables AGE18, UNDR18, and OVER18 were grouped under one variable entitled AGELEVEL. The variables NOHSG, GED, HSGRD, SOMCOL, and COLGRD were grouped under one variable entitled EDLEVEL. The variables MARRY, DEPEND1, and DEPEND2 were grouped under one variable entitled NUMOFDEP. The variables AFPTGP1, AFPTGP2, and AFPTGP3 were grouped under one variable entitled TESTSCORELEVEL. Details of the exact coding of the new grouped variables are listed from Table 4-7 through Table 4-9.

Table 4-7. Grouped Variable: RACE, Components

Individual Variables:		Grouped Variable:
		RACE
WHITE=1	→	1
BLACK=1	→	2
MINORITY=1	→	3

Table 4-8. Grouped Variable: AGELEVEL, Components

Individual Variables:		Grouped Variable:
		AGELEVEL
AGE18=1	→	1
UNDR18=1	→	1
OVER18=1	→	2

Table 4-9. Grouped Variable: TESTSCORELEVEL, Components

Individual Variables:		Grouped Variable:
		TESTSCORELEVEL
AFQTGP1,2,3=0	→	0
AFQTGP1=1	→	1
AFQTGP2=1	→	2
AFQTGP3=1	→	3

Table 4-10. Grouped Variable: EDLEVEL, Components

Individual Variables:		Grouped Variable:
		EDLEVEL
NOHSG=1	→	1
GED=1	→	2
HSGRD=1	→	2
SOMCOL=1	→	3
COLGRD=1	→	4

The logic used in Table 4-10 was to identify the different education levels and group according to similarities. GED and HSGRD were grouped based on relative equivalency while SOMCOL and COLGRD were each given a separate group identifier based on the difference between having received college level knowledge and actually graduating from college.

Table 4-11. Grouped Variable: NUMOFDEP, Components

Individual Variables:							Grouped Variable:
							NUMOFDEP
If	MARRY=0	&	DEPEND1=0	&	DEPEND2=0	→	0
If	MARRY=0	&	DEPEND1=1	&	DEPEND2=0	→	1
If	MARRY=0	&	DEPEND1=0	&	DEPEND2=1	→	2
If	MARRY=1	&	DEPEND1=0	&	DEPEND2=1	→	3
If	MARRY=1	&	DEPEND1=1	&	DEPEND2=0	→	4
If	MARRY=1	&	DEPEND1=0	&	DEPEND2=0	→	5

The logic used in Table 4-11 was to identify the different levels of dependents. As shown, each level was given a different identifier. It was felt that having one dependent versus two or more needed to be categorized separately. While a spouse is a dependent, it was essential to distinguish between marital status as well.

Grouped Variable Analysis

The following tables show output from the discriminant analysis procedures using grouped variables. For each of the three datasets, First-term Airmen, Second-Term Airmen, and Career-Term Airmen, two procedures were performed. The first procedure uses SAS system default prior probabilities of 50/50. In order to capture the uniqueness of each of the datasets the discriminant analysis procedures were ran again using prior probabilities proportional to the two populations: Stay and Leave.

Table 4-12. APER for First-Term Airmen Using Grouped Variables

Variables used: RACE, AGELEVEL, EDLEVEL, NUMOFDEP, TESTSCORELEVEL					
	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
Prior probabilities	classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
50/50	86266	45603	182854	110068	0.45018
62/38	324	109793	182854	110068	0.37593

Table 4-13. APER for Second-Term Airmen Using Grouped Variables

Variables used: RACE, AGELEVEL, EDLEVEL, NUMOFDEP, TESTSCORELEVEL					
	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
Prior probabilities	classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
50/50	61246	13043	124053	29677	0.48324
81/19	0	29677	124053	29677	0.19305

Table 4-14. APER for Career-Term Airmen Using Grouped Variables

Variables used: RACE, AGELEVEL, EDLEVEL, NUMOFDEP, TESTSCORELEVEL					
	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
Prior probabilities	classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
50/50	74233	4030	162709	9540	0.45436
94/6	0	9540	162709	9540	0.05538

As shown in the tables presented, the use of the prior probabilities shows a significant improvement in the APER. However, we see that in order to minimize error, the software program tended to classify the majority of the data as a STAY. In the case of Second-Term and Career-Term Airmen, every record was classified as a STAY. Although these results look promising it is evident that bias exists.

BY Variable

After having ran a long series of analysis with the datasets it was hoped to have at this point developed a process that would efficiently provide meaningful output. All of the previous analysis was reviewed thoroughly and a determination was made as to an appropriate route for running further analysis. Since the majority of the independent variables were successfully grouped, future discriminant analysis would be performed using solely these five variables: RACE, EDLEVEL, AGELEVEL, NUMOFDEP, AND TESTSCORELEVEL. Additionally, we decided to concentrate our analysis on a particular variable to obtain separate analyses on observations in groups. The variable in particular was AFSC. A discriminant analysis procedure will be run on each dataset using the five variables, discriminating "by" AFSC. This will allow us to identify any trends in the data based on AFSC to see if a particular AFSC is experiencing significant losses or even if a particular AFSC is experiencing a significant amount of Airmen who tend to stay on active duty.

Analysis by AFSC

The following tables show output from the discriminant analysis procedures using grouped variables by AFSC. For each of the three datasets, First-term Airmen, Second-Term Airmen, and Career-Term Airmen, a discriminant analysis procedure was performed using prior probabilities proportional to the two populations: Stay and Leave. Each dataset contained over 200 different AFSCs; the ones displayed in the tables are only a partial listing and were selected at random for illustration purposes.

Table 4-15. APER for First-Term Airmen Using Grouped Variables by AFSC

Trial #	AFSC	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
		classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
1	122	0	450	943	450	0.32304
2	1C0	3	169	381	170	0.31216
3	251	21	416	792	436	0.35586
4	326	38	786	1211	832	0.40333
5	411	72	678	1025	750	0.42254
6	566	0	411	656	411	0.38519
7	623	46	481	976	522	0.35180
8	732	53	1081	3177	1101	0.26508
9	903	29	361	402	368	0.50649
10	981	49	695	1197	700	0.39220

Table 4-16. APER for Second-Term Airmen Using Grouped Variables by AFSC

Trial #	AFSC	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
		classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
1	114	0	71	554	71	0.11360
2	208	0	255	656	255	0.27991
3	2A5	0	437	1591	437	0.21548
4	2S0	0	338	1522	338	0.18172
5	306	0	228	909	228	0.20053
6	452	0	628	3008	628	0.17272
7	571	0	256	1178	256	0.17852
8	631	0	244	1325	244	0.15551
9	924	0	152	398	152	0.27636
10	981	0	145	607	145	0.19282

Table 4-17. APER for Career-Term Airmen Using Grouped Variables by AFSC

Trial #	AFSC	0 missclass	1 missclass	n0=# of actual stays	n1=#of actual leaves	APER
		classified as a Leave but really is a Stay	classified as a Stay but really is a Leave			
1	201	0	17	476	17	0.03448
2	207	0	22	560	22	0.03780
3	208	0	44	567	44	0.07201
4	272	0	41	1230	41	0.03226
5	305	0	23	672	23	0.03309
6	306	0	38	923	38	0.03954
7	3M0	0	57	892	57	0.06006
8	426	0	40	1374	40	0.02829
9	571	0	81	1049	81	0.07168
10	981	0	34	545	34	0.05872

The results of the “Analysis by AFSC” mirror that of the “Grouped Variable Analysis.” The discriminant analysis procedure again forces classification heavily to the STAY category. One of the obstacles of analyzing the data by AFSC was that in order to perform the discriminant analysis procedure there must exist two groups. In the case of many of the AFSCs there were only one group, meaning that members of a particular

AFSC either all were listed as a STAY or all were listed as a LEAVE. This is due in part by the fact that there were AFSCs with only a few observations and many with only one observation.

Summary

This chapter contained numerous tables displaying results of the analysis performed in this study. Insight achieved through the application of the discriminant analysis technique on the demographic variables of all three datasets will be of great benefit for the understanding how to better model the Airman separation decision. The results stated in this chapter support the fact that use of the discriminant analysis procedure did not provide a meaningful way to predict enlisted separation behavior using demographic variables. In order to model the enlisted separation decision other methods are necessary. Chapter five will provide recommendations and conclude the research.

5. Recommendations and Conclusion

Outcome

The research shows that the use of the discriminant analysis data mining technique for analyzing the data specified does not give results that are useful when trying to determine Enlisted Stay/Leave populations. The results showed that although the estimates were consistent, they were severely optimistically biased. This bias favored classification into the STAY category due to the use of proportions used for prior probabilities.

Another obstacle was that due to the nature of the data, zero-one entries, in many of the cases the within-class and pooled covariance matrices were singular. The discriminant function that would have been used to classify future observations utilizes the inverse of either the two with-in class covariance matrices or the pooled covariance matrix. Since many of these were singular, hence non-invertible, there were no discriminant functions provided for these analysis runs.

An additional outcome of this research was not so much what was found but what was not found. In all of the discriminant analysis preformed during this study there appeared to be no significant evidence of trends relating to the separation decision and demographics that would give a negative image of Air Force retention patterns. By this we mean that, for example, we did not discover any increase in the number of enlisted separating based on RACE. This substantiates the notion that the Air Force acts equally and fairly to all Airmen regardless of their demographics and this is not a source of loss of personnel. In the majority of the cases marital status, number of dependents, pay grade, and education level were shown to play the most significant role in the separation decision. This was true for all three datasets. The degree to which these factors were significant is still left to be determined.

Recommendations for Future Research

This thesis was successful in analyzing the demographic data using discriminant analysis, however, use of other data mining techniques may prove to provide more useful output and results. A future researcher could possibly find a data mining technique that would be more conducive to the type of data in the files provided. Suggested techniques include Neural Networks and Factor Analysis.

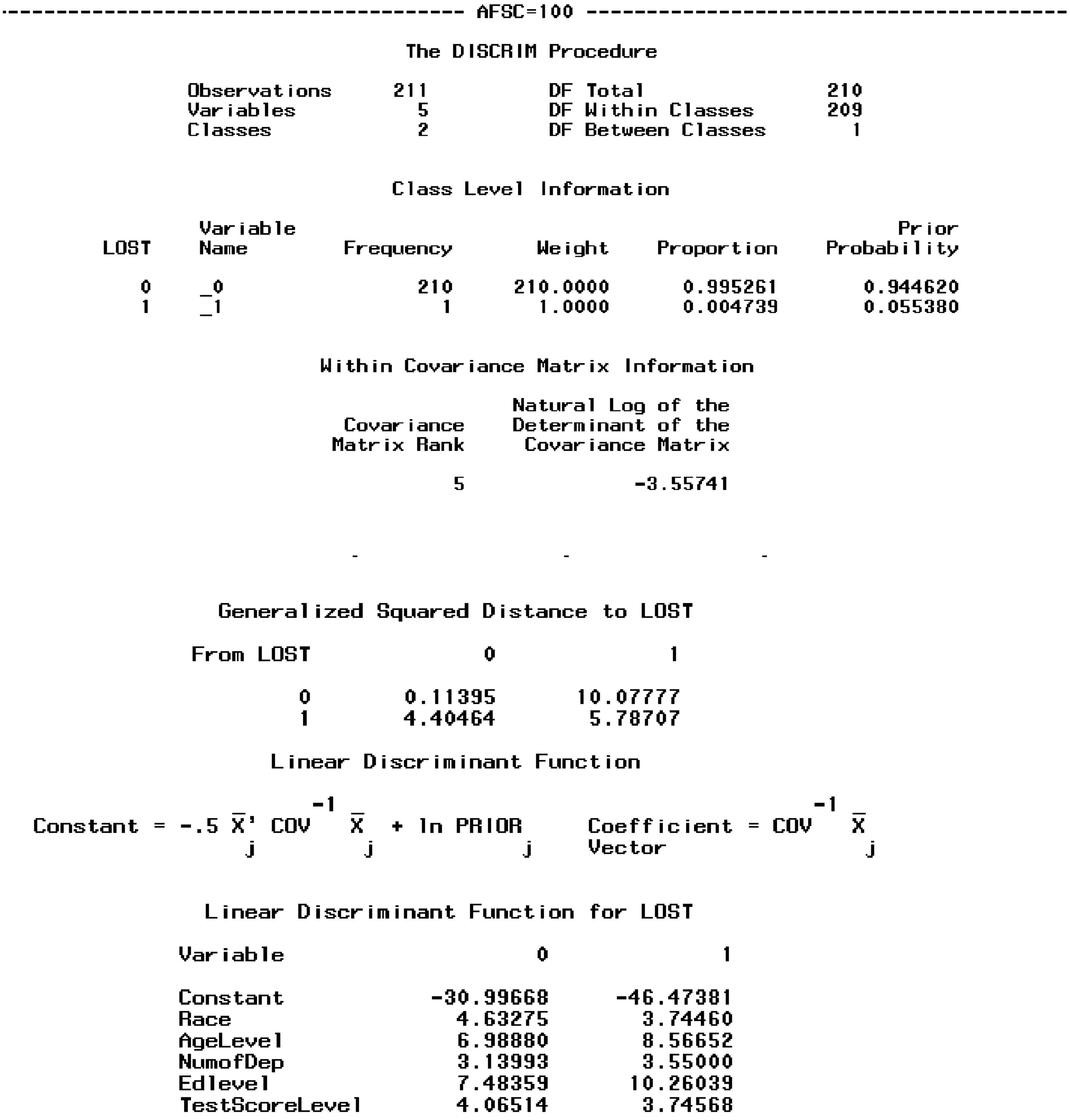
Another approach to a follow-on research effort would be to look at using discriminant analysis using demographic and economic data. One major downfall of the data used in this study was the lack of continuous or non-zero-one variables. If it were somehow able to obtain more continuous data this technique would prove more beneficial.

The best way to analyze the separation decision would be to ask those leaving active duty what factors play a major role in their decision. This is suggestive of analysis of "Exit Survey" responses. It would be an interesting study to see development and analysis of a survey given to Airmen upon separation. This survey would have to be developed in such a way as to provide variable types and responses for use in an appropriate model.

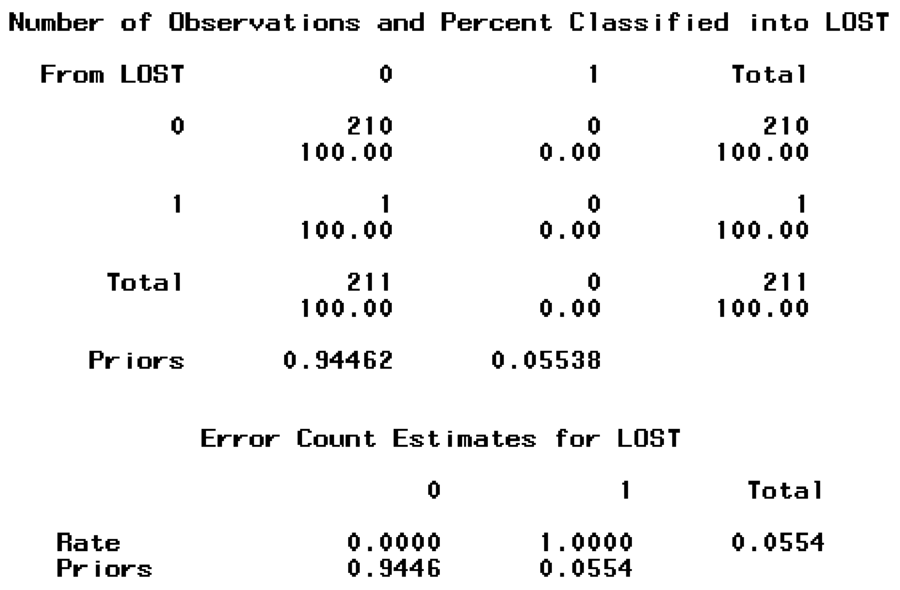
Conclusion

In conclusion this thesis analyzed demographic data using the discriminant analysis technique. This data mining technique was performed with the SAS software program. The results show that demographic variables are significant in the separation decision. However, in order to fully model the separation decision, one must consider more than just the demographic variables. In order for the Air Force to deal with the issue of retention when it comes to Airmen separating, it is imperative that demographic data are not the sole input for the model. Demographics play a role in the decision but not without other factors being considered.

Appendix: SAS Sample Output



```
------------------------------------- AFSC=100 ---------------------------------------

                                The DISCRIM Procedure

               Observations      211          DF Total                 210
               Variables           5          DF Within Classes        209
               Classes             2          DF Between Classes         1

                               Class Level Information

                Variable                                                       Prior
        LOST    Name          Frequency        Weight     Proportion     Probability

           0    _0                  210      210.0000       0.995261        0.944620
           1    _1                    1        1.0000       0.004739        0.055380

                         Within Covariance Matrix Information

                                          Natural Log of the
                              Covariance  Determinant of the
                             Matrix Rank   Covariance Matrix

                                       5            -3.55741



                             -                 -                -


                    Generalized Squared Distance to LOST

                  From LOST            0              1

                          0      0.11395       10.07777
                          1      4.40464        5.78707

                        Linear Discriminant Function

                             -1  _                                              -1  _
      Constant = -.5 X' COV      X  + ln PRIOR      Coefficient = COV    X
                     j           j            j     Vector                          j

                    Linear Discriminant Function for LOST

              Variable                           0              1

              Constant                   -30.99668      -46.47381
              Race                         4.63275        3.74460
              AgeLevel                     6.98880        8.56652
              NumofDep                     3.13993        3.55000
              Edlevel                      7.48359       10.26039
              TestScoreLevel               4.06514        3.74568
```



```
Number of Observations and Percent Classified into LOST

  From LOST            0            1         Total

          0          210            0           210
                  100.00         0.00        100.00

          1            1            0             1
                  100.00         0.00        100.00

      Total          211            0           211
                  100.00         0.00        100.00

     Priors      0.94462      0.05538


            Error Count Estimates for LOST

                          0            1        Total

  Rate               0.0000       1.0000       0.0554
  Priors             0.9446       0.0554
```

Bibliography

Andrews, D. F., R. Gnanadesikan, and J. L. Warner. "Methods for Assessing Multivariate Normality," Multivariate Analysis-III "Proceedings of the Third International Symposium on Multivariate Analysis held at Wright State University, Dayton OH, 19-24 June 1972". Academic Press. New York, 1973.

Basalla, Mark A. A Methodology for the Analysis and Prediction of Air Force Officer Retention Rates. MS Thesis, AFIT/GOR/ENC/96M-01. Graduate School of Engineering of the Air Force Institute of Technology, Wright-Patterson AFB OH, March 1996.

Bauer, K. W. "OPER685, Applied Multivariate Data Analysis," Fall 2001. Air Force Institute of Technology, Wright-Patterson AFB OH.

Callander, Bruce D., "Drawdown and Pain," Air Force Magazine, 75, 38-41 (January 1992).

Callander, Bruce D., "Going: A Fifth of the Force," Air Force Magazine, 74, 36-39 (February 1991).

Chernoff, Herman. " Some Measures for Discriminating between Normal Multivariate Distributions with Unequal Covariance Matrices," Multivariate Analysis-III "Proceedings of the Third International Symposium on Multivariate Analysis held at Wright State University, Dayton OH, 19-24 June 1972". Academic Press. New York, 1973.

Correll, John T., "Pain and Regeneration," Air Force Magazine, 75, 4 (April 1992).

Dillon, William R., and Matthew Goldstein, Multivariate Analysis, Methods and Applications. John Wiley & Sons. New York, 1984.

Directorate for Information Operations and Reports (DIOR). "Military Personnel Statistics". Jan 2001, http://web1.whs.osd.mil/mmid/military/miltop.htm.

Fayyad, Usama M., Gregory Piatetsky-Shapiro , Padhr Smyth, and Ramasamy Uthurusamy, Advances in Knowledge Discovery and Data Mining. Massachusetts: The MIT Press, 1996.

General Accounting Office (GAO). Military compensation: Selected Occupational Comparisons with Civilian Compensation. GAO/NSIAD-86-113. Washington: General Accounting Office, June 1986.

Gill, H. Leroy and Donald R. Haurin. "Staying in the Military and Spouse Earning Opportunities," Unpublished paper. Air Force Institute of Technology, Wright-Patterson AFB OH, June 1992.

Giri, Narayan C. Multivariate Statistical Analysis. Marcel Dekker, Inc. New York, 1996.

"How Some Companies Hold Turnover to 10% or Less," HRfocus , S1-S3, (July 2000).

Hutchins, Jennifer. "Getting To Know You," Workforce, 44-48 (November 2000).

Kiger, Patrick. "Retention On The Brink," Workforce, 58-65 (November 2000).

Lommen, Peter. A Methodology For the Analysis and Prediction of Air Force Enlisted Aircraft maintenance. MS Thesis, AFIT/GLM/LAL/99-S-08. Graduate School of Engineering and Management of the Air Force Institute of Technology, Wright-Patterson AFB OH, September 1999.

Longhorn, David C. Using Simulation to Model an Army recruiting Station with Seasonality Effects. MS Thesis, AFIT/GOR/ENS/00M-18. Graduate School of Engineering of the Air Force Institute of Technology, Wright-Patterson AFB OH, March 2000.

Mehay, Stephen L. "Analysis of Performance Data for Junior Navy and Marine Corps Officers," Naval Postgraduate School, Monterey CA, October 1995.

Moore, Carol S., Henry S. Griffis and Linda C. Cavalluzzo. "A Predictive Model of Navy Second-Term Retention," Center for Naval Analyses, Alexandria VA, April, 1996.

Nakada, Michael K. and James P. Boyle. "Nuclear Officer Retention: An Economic Model," Navy Personnel Research and Development Center, San Diego CA, March, 1996.

National Defense Authorization Act for Fiscal Year 2000, title VI (P.L. 106-165, October, 1999).

Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. Mathematical Statistics with Applications, Duxbury Press, New York, 1996.

Williams, Jason. Idendifying Demand Indicators For Air Force Recruiting Service With Discriminant Analysis. MS Thesis, AFIT/GOR/ENS/01M-18. Graduate School of Logistics and Acquisition Management of the Air Force Institute of Technology, Wright-Patterson AFB OH, March 2001.

Withers, Pam. "Retention Strategies That Respond to Worker Values," 36-48 Workforce, (July 2001).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)* 20-03-2002	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From – To)* July 2001 – Mar 2002

4. TITLE AND SUBTITLE

IDENTIFYING ENLISTED STAY AND LEAVE POPULATION CHARACTERISTICS WITH DISCRIMINANT ANALYSIS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hoggard, Zabrina, Y., 1st Lieutenant

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/02-08

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

Steven L. Forsythe, Major, USAF
Senior Enlisted Personnel Analyst
AFPOA
1235 Jefferson Davis Hwy, Suite 301
Arlington, VA 22202-3283
DSN:664-0767

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

There exist factors that play a major role in an enlisted Airman's decision to either stay on active duty in the Air Force or separate. The current force structure of the U.S. Air Force and increased loss of enlisted personnel is a major concern as we look at maintaining manpower to meet the needs of the Air Force. The Air Force is reacting to this low retention problem by increasing the bonuses for initial enlistments and reenlistments, home basing, increasing quality of life for Air Force personnel with enlisted dormitory plus-ups, and under AEF personnel have increased predictability of deployment.

This thesis provides a method for identifying the variables that most characterize Stay and Leave populations for enlisted Airmen on active duty in the Air Force. Discriminant Analysis is used to identify population characteristics that categorize the two groups. A methodology is constructed that can discriminate between Airmen that stay on active duty military service and Airmen that leave active duty military service.

15. SUBJECT TERMS

Discriminant Analysis, Enlisted Retention

16. SECURITY CLASSIFICATION OF: UNCLASSIFIED			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	b. ABSTRACT	c. THIS PAGE			John O. Miller, Lt Col, USAF
U	U	U	UU	65	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4326 (John.Miller@afit.edu)