3-1998

A Model to Forecast Civilian Personnel Inventory for the National Security Agency

Stephen G. Hoffman

A MODEL TO FORECAST CIVILIAN PERSONNEL INVENTORY FOR THE NATIONAL SECURITY AGENCY

THESIS

Stephen G. Hoffman
Captain USAF

AFIT/GOR/ENS/98M-13

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GOR/ENS/98M-13

A MODEL TO FORECAST CIVILIAN PERSONNEL INVENTORY FOR THE NATIONAL SECURITY AGENCY

THESIS

Stephen G. Hoffman
Captain USAF

AFIT/GOR/ENS/98M-13

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

AFIT/GOR/ENS/98M-13

A MODEL TO FORECAST CIVILIAN PERSONNEL INVENTORY FOR THE NATIONAL SECURITY AGENCY

THESIS

Presented to the Faculty of the Graduate School of
Engineering of the 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

Stephen G. Hoffman, B.S., M.P.A
Captain USAF

March 1998

THESIS APPROVAL

Student: Stephen G. Hoffman, Capt, USAF Class: GOR-98M

Title: A Model to Forecast Civilian Personnel Inventory for the National Security Agency

Defense Date: 25 February 1998

Committee: Name/Title/Department Signature

Advisor Kenneth W. Bauer, Jr.
Professor
Department of Operational Sciences

Reader Edward F. Mykytka
Head
Department of Operational Sciences

Acknowledgments

Completing the class work and thesis required for graduation have been very much a team effort and I am grateful to all of my teammates. I would like to thank Sue Rose and Mary Filippell of the National Security Agency's Workforce Assessment Team for their promptness in collecting data as well as responding to my endless stream of questions.

Thanks also to my advisors Dr. Kenneth Bauer and Dr. Edward Mykytka for their guidance - even though I didn't like hearing it at the time. Their suggestions and encouragement helped me to stay on that straight and narrow way.

I appreciate Bryan Livergood and Eric Beene, who proved their friendship by not only sticking with me on class projects, but also by enduring all of my bad jokes.

Finally and foremost, I would like to thank my wife and kids, Karen, Jessica, and Joshua, for continually supporting me through this program. In particular Karen, whose faith reassured me when I had only uncertainties.

It is a shame that only one name will appear on the diploma!

Stephen G. Hoffman

Table of Contents

List of Figures

List of Tables

Abstract

The purpose of this thesis is to develop a model to forecast civilian personnel inventory for the National Security Agency (NSA). Accurate prediction of personnel inventories will help the NSA with issues ranging from hiring, promotion, and ethnic/gender diversity to the development of cost-effective force reduction programs. The model must be capable of producing multiple-year forecasts of the NSA's civilian force structure with a minimum fidelity level that allows for break-out by years-of-service, career field, and grade. Historical data will be utilized to estimate logistic regression parameters that will compose the internal workings of the model. The JMP statistical software package will be used for data analysis, while the model will be developed as a Microsoft Excel application.

A MODEL TO FORECAST CIVILIAN PERSONNEL INVENTORY FOR THE NATIONAL SECURITY AGENCY

I. Introduction

The Downsizing Federal Government

The Federal Government is downsizing and decision makers at all levels of the bureaucracy are feeling pressure to do more with less. The downward trend at the Federal Government level has forced Department of Defense (DoD) officials to utilize the civilian workforce as the primary vehicle for future employment reductions. Civilian human resource division managers who have watched the military reduce manning by a third over the last ten years are now being asked to produce similar percentages of additional losses.

Total active duty military positions from 1989 to 1997 have dropped from 2,130,000 to 1,450,000 and are expected to decline to 1,360,000 by 2003 (Quadrennial Defense Review 1996). While these future losses are substantial, the largest proportion of active duty military reductions has already occurred. The Quadrennial Defense Review (QDR) indicates that future reductions in the DoD will be primarily achieved through the civil service.

The QDR contains forecasted personnel data through fiscal year 2003. The baseline forecast contains 80,000 more positions than the QDR goal of 640,000 DoD civil servants. The baseline predictions, which come from the personnel analysts at the Pentagon, are based upon the existing body of personnel policy. Table 1 compares the historical civilian employment trends of the entire Federal Government with those of the DoD. Also shown are QDR forecasted estimates of future DoD civilian employment.

Table 1. Federal Civilian Employment

YEAR	FEDERAL	DOD
1989	3,123,731	1,075,437
1990	3,128,267	1,034,152
1991	3,111,912	1,012,715
1992	3,085,323	982,774
1993	3,013,508	921,179
1994	2,971,584	879,878
1995	2,918,674	830,738
Fcst FY1997		800,000
Fcst FY2003		720,000
Goal FY2003		640,000

(Historical data from 114th & 116th Editions of the Statistical Abstract of the United States; Forecasted DoD data from QDR 1996).

The portion of the DoD which is the focus of this thesis is the National Security Agency (NSA). The impending reductions in the NSA, as part of the Federal Government and the DoD's continued draw-down, indicate a need for personnel

programs which will not only meet the required downsizing goals, but more importantly, will retain the NSA's functionality. The method as well as impetus for attaining this reduced Government size comes from the highest levels as noted by President Clinton's April 4, 1995 comments:

> To reduce the work force by 102,000 positions by the end of fiscal 1994, we offered about 70,000 buyouts. Several non-DoD agencies have offered deferred buyouts that will take place between now and March 1997. Defense will be using buyouts as it continues to downsize through 1999. Counting those, we expect to buy out another 84,000 workers through 1997 as we work to reduce the work force by a total of 272,900 positions (Clinton 1995).

In Section V (Forces and Manpower), the QDR states, "The issue is not whether we will reshape our forces, but how and when" (Quadrennial Defense Review 1996). The DoD is seeking additional buyout authorities to assist in the process of reaching the goals mentioned earlier. As part of the DoD, the NSA anticipates that it will be tasked to produce additional losses in an effort to help the DoD and Federal Government meet their respective civilian work force goals. Accepting this responsibility and being able to predict the resulting consequences are two different things. For this reason the NSA has made efforts to better understand the interdependence of its personnel policies and programs. The model developed in this thesis will assist in the NSA's efforts to both understand and improve its human resource activities.

Problem Statement

The NSA desires the ability to forecast its civilian personnel force structure for multiple fiscal years. The predictions will be routinely updated and used to aid in the management of numerous human resource areas such as new hires, promotions, draw-down policy, and ethnic/gender diversification. Personnel data is available for eight historical years from three different source files. The current "snapshot" of the civilian NSA inventory is contained in the EDPSMSTR file, which is updated monthly. Eight years of end-of-fiscal-year EDPSMSTR files are archived as PASTEDPS files. The PASTEDPS records are historical text files that contain the same fields as the current EDPSMSTR file. Also available is data from the ACTIVITY file which contains coded data fields representing personnel actions and their respective action dates.

Objective

The objective of this thesis is to develop a model which will intelligently use both historical data and the NSA's human resource expertise to accurately forecast civilian personnel inventory for multiple years. These forecasts must have a level of detail sufficient to meet the needs of the NSA's Workforce Assessment Team. As a minimum, the forecasts must allow for break-out by years of service,

career field, and grade. The previously mentioned historical databases contain 86 common fields for each individual. The fields associated with a social security number represent an individual's personal attributes (age, grade, race, etc.) which the NSA tracks. My initial screening yielded 15 fields with potential utility in the modelling effort. Most of the fields screened out were strictly for the NSA's in-house use. Some examples are FLSA (Fair Labor Standards Act Code), HOURS (hours per week), and SSN (Social Security Number). Brief descriptions of the fields that I initially requested from the NSA are contained in Table 2.

Table 2. Fields Contained in Initial Data Set

FIELD	DEFINITION
ACT_CODE	Action Code
ACT_DTE	Action Date
COSC	Career Occupation Specialty Code
DOB	Date of Birth
DOG	Date of Grade
EDLVL	Education Level
EMPCAT	Category of Employment
EOD	Date of Service Entry (Federal)
FUNCTL	Functional Category Code
GRD	Grade
LOCUMSAL	Total Salary
RSC	Recruitment Source
SCDR	Date of Service Entry (NSA)
SRC	Sex Race Code
VP	Veteran's Preference Code

(Supplied by NSA, 1997)

Major Personnel Functions

There are many personnel functions which civilians at the NSA may experience during any given year. For example, at any point during the year a person may get promoted or demoted. Another possibility is that the person's career field may change due to a voluntary lateral move, or possibly a mandatory management decision. This movement of people from one career field to another is called *migration*. Any lateral move to a non-NSA agency is a *loss* not a migration. Sometimes people will migrate in conjunction with a promotion. Each year many people choose to leave the NSA prior to retirement. These non-retirement losses, where no incentive is provided, are called *attrition*. All NSA civilians have the option to either attrit or retire (if eligible), provided they are not currently repaying a service commitment obligation. Those who choose attrition or retirement are called *basic losses*. Also, many people have the option of accepting cash incentives or early retirements to leave during the year. Those who choose to accept these offers to leave are called *policy losses*.

Since all of these personnel functions tend to be interdependent, accomplishing the goal of forecasting civilian personnel for several years requires the simultaneous consideration of all five. Limitations in the

historical data (as well as common sense) dictate that only the major personnel functions should be addressed. The five major personnel functions that I have chosen to model are; basic losses, policy losses, migration, promotion/demotion, and new hires. Figure 1 shows the seven stages which represent the flow of civilians through the major personnel functions.

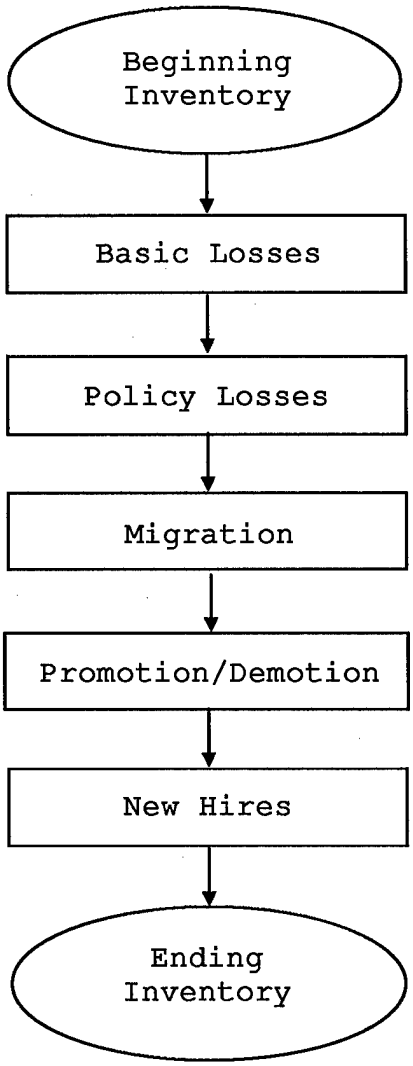


Figure 1. Flow Chart for Major Personnel Functions

Each of the seven component stages shown in Figure 1 could, in their own right, be modelled independently. If the goal were simply to predict basic losses, then only that one personnel function would need to be modelled. However, since the NSA desires the ability to forecast the entire civilian force structure, all seven stages must be assimilated into a single model. The additional requirement of multiple-year forecasts prompted me to design the model as a single-year forecasting tool which could be executed iteratively for as many years as necessary.

The model will be a personal computer application that will allow the customer (NSA's Workforce Assessment Team) the ability to set "control levers." These control lever settings drive aggregate levels of basic losses, policy losses, and new hires. Control levers will be particularly useful when the customer has specific information about how many losses or hires will occur in a particular future year. Consider the following hypothetical scenario:

> Information channeled down from an organization higher in the Federal Government has indicated to the NSA's Workforce Assessment Team that 500 additional losses must occur in each of the next three fiscal years. In order to estimate the mid-range impact of these additional losses a five-year forecast of civilian personnel inventory is desired. The appropriate setting for the policy loss control lever for each of the first three years will provide input to the model described previously. Using this customer input the model can be utilized to produce the five-year forecast of the NSA's civilian personnel inventory.

Merging each of the five personnel functions into a single application provides the customer with an easy-to-use civilian inventory forecasting tool. Single-year prediction intervals offer the customer an option to iteratively produce multiple-year forecasts. The customer control lever inputs are available to the NSA's Workforce Assessment Team to increase the model's utility. I have named the application, which contains all seven stages shown in Figure 1, the Iterative Civilian Forecast Model (ICFM).

Process Overview

Using the ICFM to predict a multiple-year civilian personnel inventory is an iterative task in which human resource managers may use individual and corporate knowledge of what has happened recently in conjunction with anticipated policy changes to predict a civilian force structure of the future. Customer decisions and inputs drive the process of creating forecasted inventories for the desired length of time. Figure 2 depicts a schematic of the process required to utilize the ICFM for prediction of multiple-year civilian personnel inventories.

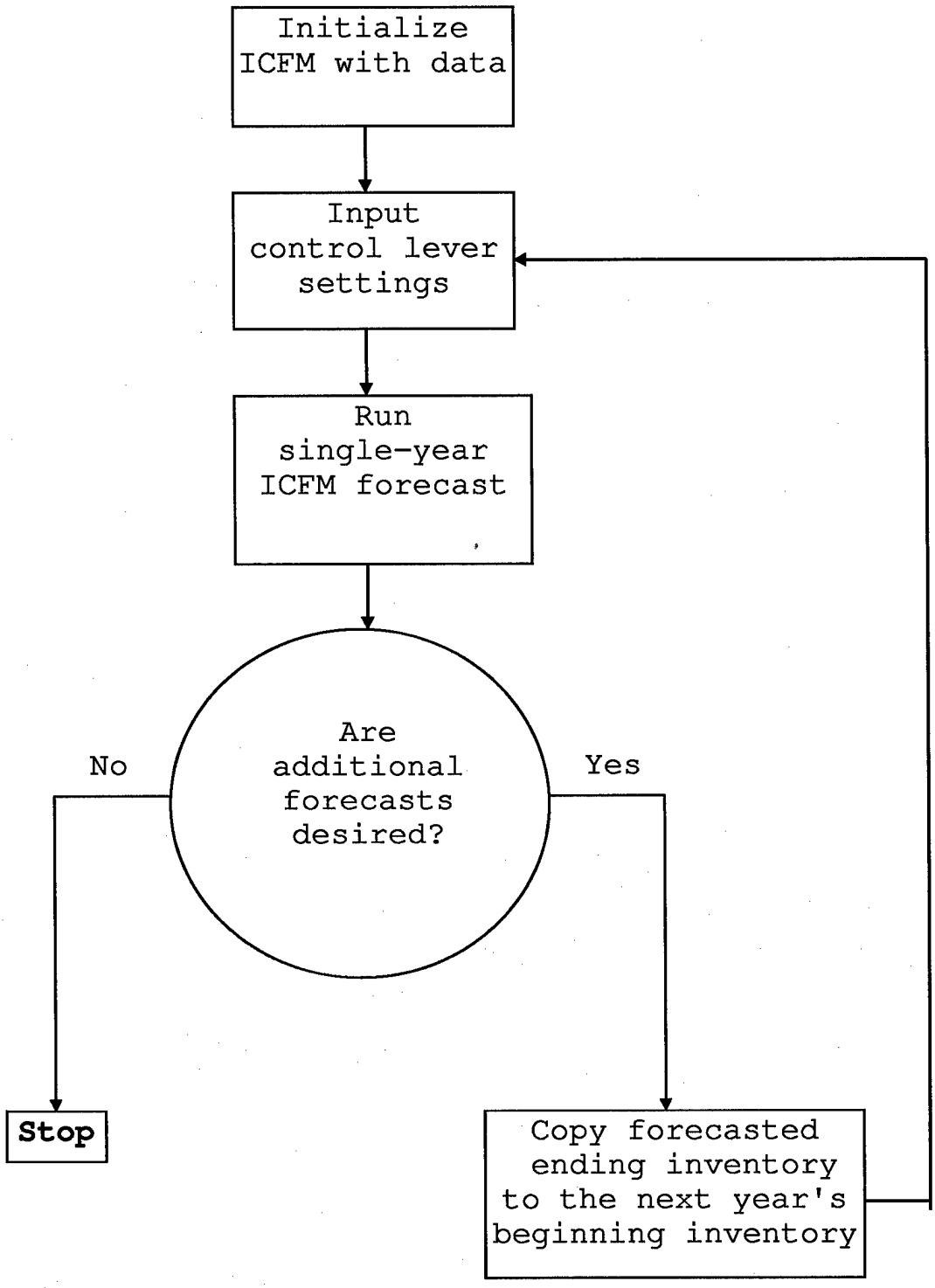


Figure 2. Decision Tree for Multiple-Year Forecast

The ICFM only requires a few customer inputs to produce a forecasted inventory. The initial requirement is a beginning inventory extracted from the NSA's personnel database. The current fiscal year's beginning inventory is always contained in the previous 30 September's EDPSMSTR file. The personnel data in this particular snapshot characterizes the NSA's civilian force just one day prior to the beginning of the current fiscal year. Since historical data is archived on a fiscal year basis, the predictions are constrained to one-year increments. The remaining customer inputs are contained in three control lever settings. These settings control the aggregate amount of basic losses, policy losses, and new hires that the ICFM will produce en route to a single-year civilian personnel inventory forecast.

After the ICFM has produced the single-year forecast, the operator must decide to either stop, or iterate through the forecasting process again. This decision is completely dependent upon the required number of forecast years. If the decision is to produce an additional year's forecast, then the current year's forecasted ending inventory is copied to the next year's beginning inventory so the ICFM can produce an additional single-year civilian personnel inventory forecast. If at any point, the original single-

year output is deemed unacceptable, customer control levers may be adjusted. Changing the control lever settings immediately alters the results for the single-year currently being forecasted. By modifying the control lever settings, the customer can quickly and easily determine the sensitivity of the forecast to basic losses, policy losses, or new hires.

The first and most important personnel function to be modelled is basic losses. Action codes in the ACTIVITY file are used to group losses into either the basic or policy type. Action codes which represent retirement, resignation, death, or transfer to non-NSA positions are categorized as basic losses. These basic losses are the primary contributing factor to the internal condition of the civilian force. Since the fidelity level of the forecasts required by the NSA includes (among others) years-of-service, career field, and grade, the effort to model all personnel functions from basic losses to new hires must be accomplished at that same level. To assure common dimensionality, the data set received from the NSA was converted into useable variables. The 15 fields that contained the civilian personnel information in the original data set received from the NSA were used to create the 13 variables described in Table 3. These 13 variables are used throughout the ICFM.

Table 3. Description of Variables Created From NSA Data

VARIABLE	TYPE	DEFINITION
Bsc_Loss	Binary	Indicates Basic Loss
Plcy_Lss	Binary	Indicates Policy Loss
ECF	Nominal	End-of-Year Career Field
EGRADE	Nominal	End-of-Year Grade
New_Hire	Binary	Indicates New Hire
AGE	Continuous	Age
YOS	Continuous	Years-of-NSA-Service
TIG	Continuous	Time-in-Grade
BGRADE	Nominal	Beginning-of-Year Grade
GENDER	Nominal	Gender
RACE	Nominal	Race
BCF	Nominal	Beginning-of-Year Career Field
EDCAT	Nominal	Category of Education

After basic losses, the next personnel function modelled in the ICFM is policy losses. Policy losses are all losses which came as a result of a draw-down, buyout, early retirement, or other specific program intended to encourage NSA civilians to voluntarily terminate employment. Several programs have been offered in recent years designed to prompt additional losses as the NSA attempts to reduce its manning. Modelling the different programs is problematic because many of them target a specific subset of employees. Adding to the difficulty of distinct models for each program is the fact that the rules to these programs sometimes varied over time. For this reason, all programs intended to produce losses over-and-above normal attrition and retirement losses are grouped into a single policy loss category.

Policy losses in the ICFM will be highly dependent upon customer input at the aggregate level. While policy losses are not the only personnel function for which the customer inputs a control lever setting, it is the most volatile due to varying draw-down requirements. From one year to the next, these politically driven draw-down requirements may cause policy losses to fluctuate quite a bit.

The records existing after basic and policy losses have been removed from the beginning inventory are called the *survived inventory*. These first two personnel functions modelled remove all forecasted losses for a single-year forecast. The next personnel function modelled in the ICFM is the cross-flow of workers among the ten career fields. These cross-flows, called *migrations*, include lateral moves into and out of NSA positions. Lateral transfers, administrative moves or other job related movement where the individual is permanently awarded a career field different than what was held previously are included under the heading of migrations. Approximately 25% of all records received from the NSA indicated a migration. All migrations are caused by either a voluntary lateral move, mandatory managerial decision, office restructure, or in conjunction with a promotion. These four reasons for cross-flow provide a picture of the need to accurately predict migrations. On the one hand, migrations do not affect the total size of the

force so modelling them may appear unnecessary. On the other hand, the customer's desire for high resolution output dictates that correctly predicting the internal structure of the inventory is just as important as correctly predicting its aggregate size.

In the promotion/demotion stage of the ICFM, records in the *migrated inventory* are given an opportunity to increase or decrease their grade. The promotion/demotion stage could also be called "grade change" since the remaining inventory face both promotion or demotion. The vast majority of actual grade changes are promotions, some of which are coupled with a change of career field. These migration-dependent promotions are somewhat unpredictable because both the promotion opportunity as well as manning levels of all career fields are necessary to predict them. However, the vast majority of grade changes are promotions where no change of career field occurs. The method chosen to model both the promotion and migrations personnel functions are very similar. In both cases the total number of records within the ICFM does not change.

As was the case for migrations, modelling promotions/demotions in the ICFM may seem unnecessary. But, accurately predicting both migrations and promotions is critical to the credibility of a multiple-year forecast. Since the basic and policy losses are calculated at a

fidelity level which includes both career field and grade, any multiple-year forecast is dependent upon a good prediction of both. For each additional number of forecast years required by the customer, an incorrectly predicted internal structure will increasingly decrease the model's accuracy. This interdependence of major personnel functions is precisely the reason that the ICFM will be useful to the NSA's Workforce Assessment Team.

The hiring of new people is the next personnel function modelled in the ICFM. As in all of the previous stages, historical data drives this calculation. The new hires which are forecasted will be placed in career field and grade positions according to historical proportions. Additionally, the new hires will be distributed across gender and ethnicity according to historical rate. Generally, the aggregate size of the group of newly hired individuals added in the ICFM is very close to the number of records subtracted out due to basic losses. These two sections in the ICFM allow the aggregate inventory to remain roughly equal in the aggregate. The ending inventory represents an actual beginning inventory which has had basic and policy losses removed, migrations and promotions accomplished, and new hires added.

If, upon inspection of the forecasted civilian personnel inventory, the operator is unsatisfied with the

aggregate number of losses or new hires, customer control lever settings may be adjusted and the ICFM rerun. The *ending inventory* may be output in the aggregate or by function according to the customer's desire. If an additional fiscal year forecast is necessary the customer may use the ending inventory as the next forecast year's beginning inventory. All ICFM calculations are accomplished automatically, so producing a multiple-year forecast of the NSA's civilian personnel inventory is a quick and easy task. The ICFM will produce forecasts for as many years as the customer requires.

The ICFM will offer standard formatted output as well as ad hoc requests to satisfy customer data needs. Numerous output formats, including data tables or charts, will be available to assist NSA decision makers. By saving the data sets created during the modelling runs, specific data requests can be easily accomplished at a latter date. Customer control lever settings may be adjusted repeatedly to produce comparisons of competing proposals. Once all control lever settings have been finalized and ICFM runs completed, the resulting output can be used to answer a wide variety of personnel planning and programming questions.

II. Literature Review

Logistic Regression

Logistic regression is a method of modelling ordinal, nominal, or binary response data. While nominal and ordinal response variables both contain categorical data, the latter implies value or order to the variable contents while the former does not. Binary variables, which are just special cases of nominal variables, are useful for success/failure or two-category data. Since the personnel functions to be modelled in this thesis (basic losses, policy losses, migrations, promotions, and new hires) are all characterized by nominal or binary responses, logistic regression is an appropriate method of forecasting. Figure 3 shows a sample logistic regression line where the binary response is *basic loss* and the independent variable is *years-of-service*

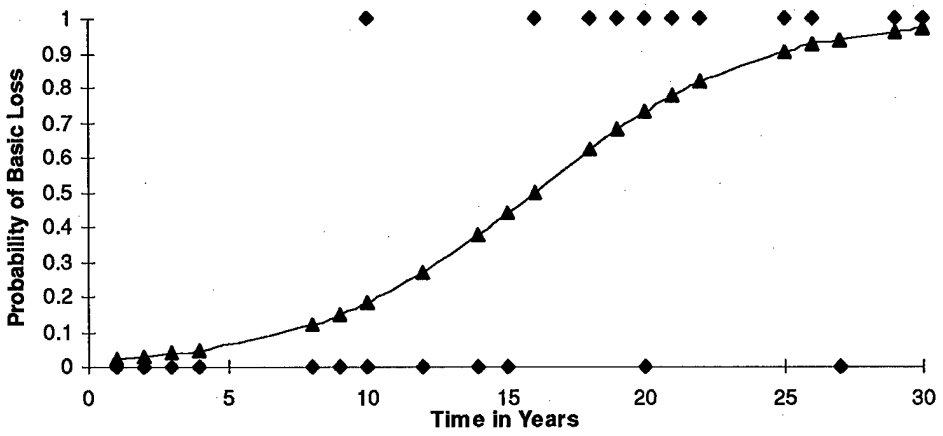


Figure 3. Logistic Regression Plot

A logit equation transforms the ordinary linear model response into a form which naturally fits binary response data (Agresti, 1990:81-94). Christensen (1990:233) describes the logit transformation as one that takes a number p between 0 and 1 and transforms it to $x = \log\left[\frac{p}{(1-p)}\right]$. The logistic transformation takes a number x on the real line and transforms it to $p = \frac{e^x}{1+e^x}$. Nominal response variables are analyzed with a simple extension of the logit model. For a binary response, a logit response model is

$$X\beta = \log\left[\frac{P(y=1)}{P(y=2)}\right]$$

which can be written $P(y=1) = F(X\beta)$ where $F()$ is the cumulative distribution function of the logistic distribution,

$$F(x) = \frac{e^x}{1+e^x} \quad .$$

The extension for r responses is to relate each response probability to the r^{th} probability, and fit a separate set of design parameters to these $r-1$ models. For a nominal response, a logit response model is

$$X\beta_{(j)} = \log\left[\frac{P(y=j)}{P(y=r)}\right] \text{ for } j=1,\ldots,r-1 \text{ (JMP 1995:545).}$$

Multivariate Techniques

Multivariate analysis techniques will be used to transform correlated variables into uncorrelated linear combinations and to investigate outliers in the data. The individual records containing personal attributes will be considered objects in the analysis. Dillon and Goldstein suggest using a "supermatrix" to visualize simultaneous relationships among the multivariate data (1984:3). Figure 4 shows an example of a supermatrix dimensioned by Time, Attribute, and Object.

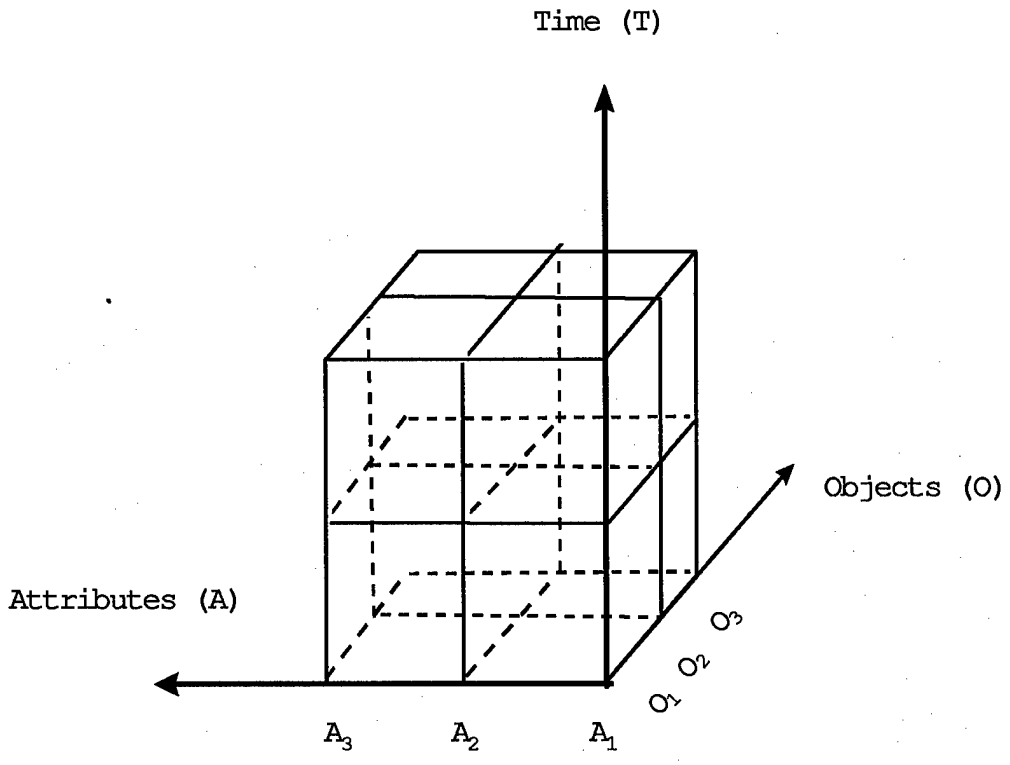


Figure 4. Supermatrix

Principal component analysis (PCA) is a data reduction technique useful in understanding data with large numbers of variables. In PCA, the original variables are transformed into a smaller set of linear combinations, called principal components, that account for most of the variation in the original set. The typical purpose of PCA is to determine the fewest number of principal components which explain the greatest proportion of the total variation. The first principal component for a data set with p original variables is:

$$PC_1 = w_{(1)1}X_1 + w_{(1)2}X_2 + \cdots + w_{(1)p}X_p$$

where the weights $w_{(1)1}, w_{(1)2}, \cdots, w_{(1)p}$ have been chosen to maximize the ratio of the variance of PC_1 to the total variation. The second principal component is uncorrelated with the first and is weighted so that it maximizes the ratio of variance of PC_2 to the total variance not already accounted for by PC_1 (Dillon and Goldstein, 1984:24). The remaining principal components are constructed in an identical fashion so they are all uncorrelated linear combinations of each other.

In any multivariate analysis, data points that are considerably different from the others have a disproportionate influence on statistical tests. All points identified as possible outliers, however, shouldn't

necessarily be removed from the database. Sometimes points that are much different from the rest actually help portray rarely occurring situations. The first step in an outlier analysis should be to identify all possible outliers and then verify those records' validity. Investigation of these outliers is especially important when building a regression model because the resulting parameter estimates will be directly affected by incorrect data. Multivariate outliers can be identified by calculating the Mahalanobis distance from each point to the multivariate mean (centroid) and then further investigating those which are significantly different from the rest. This distance formula is given by:

$$D^2 = \left(X_i - X_j\right)^T S^{-1} \left(X_i - X_j\right)$$

where S is the pooled within-group covariance matrix, and X_i and X_j are the two points for which the distance is desired. Ordinary Euclidean distances could also be utilized to identify outliers, but the Mahalanobis distance has the advantage of explicitly accounting for any correlations that might exist between the variables (Dillon and Goldstein, 1984:163).

Personnel Modelling

Forecasting civilian personnel inventories requires much more than just a model which produces reams of computer generated output. Woolsey states, "A manager would rather live with a problem that he cannot tolerate than use a solution he cannot understand" (1975:169). For this reason, the focus of this thesis is on development of a customer-oriented model which provides technically sound predications via a user-friendly application. In this section, special attention has been given to forecasting methods and existing personnel models which may help to optimize the utility of the model designed in this thesis.

Econometric models attempt to model the multiple interdependencies of sub-processes within the overall process being modelled. Makridakis et al. state that econometric models were first used in the late 1960s to forecast macroeconomic issues (1983:Ch 7). Since great detail is built into econometric models, their predictive power is considered to be very robust. Economic models may be affected by a great number of interdependent factors but a robust model is capable of accurate predictions regardless of the input conditions. This complexity is a two edged sword which makes development, operation, and maintenance a costly and time consuming endeavor. Resultantly, Makridakis et al. (1983:Ch 7) recommend that econometric models only be

implemented when the sponsoring organization is a government or extremely large corporation.

A multitude of logistic regression models exist in the literature. One example is Yan's (1992) dissertation in which an analysis of college dropouts was developed using binary response logistic regression. In her analysis, Yan attempted to determine the factors that predict voluntary withdrawal from school at any time during a students course of study (1992:62). Qualitative independent variables such as gender and ethnicity along with quantitative predictors such as age, high school GPA, and SAT score were used in the same model to predict the probability of college dropout (Yan, 1992:62). A subset containing only dropouts was collected as input to the regression model. From this data set the relative predictive power of the independent variables was determined. The model was not developed as an institutionalized tool but rather a one-time analysis of the dropout situation.

The RAND Corporation accomplished a comprehensive analysis for the Air Force in the early 1980s to develop a system of models to assist in the management of many different aspects of the enlisted personnel force structure. RAND's prototype included interaction between a system of models. This interaction provided a check and balance structure to eliminate the "stove-pipe effect" which had

initially prompted the Air Force to request the study. Prior to the RAND study the Air Force used several different models to assist in enlisted personnel policy development. Since each of the models were developed and operated independently, they often produced contrary output.

A foundational piece of RAND's Enlisted Force Management System (EFMS) was a regression model which served as a feeder for the other component models. The Bonus Effects Model (BEM) and Disaggregate Middle-Term Inventory Projection Model (DMI) both receive coefficients from the regression model. The BEM was designed to assist in the development of optimal Selective Reenlistment Bonus (SRB) plans. Selective Reenlistment Bonuses are offered to enlisted personnel in certain career fields where under-manning is an problem. The DMI was designed to assist in the development of optimal retraining plans. To meet this goal the model allows customer input and override in many personnel areas to forecast the future enlisted force structure. The BEM, DMI and other EFMS models feed each other to assure a continuity among all output products. All of the EFMS models, however, provide their respective customer with override ability (RAND, 1991:R-3600-AF).

The process of forecasting the NSA's civilian personnel inventory can benefit from positive aspects of each of these forecasting methods and personnel models. The econometric

models attempt to completely define all sub-processes resulting in a robust forecast. The required maintenance and cost of an econometric model is prohibitive in this case but their institutionalized nature appears to be valuable. Binary response logistic regression is definitely applicable to the modelling portion of the civilian inventory forecasting process. Being able to include both qualitative and quantitative predictor variables in the same regression model will certainly increase the accuracy and usability of the model. The major lesson learned from RAND Corporation's EFMS study is that customer input during design, and override ability during implementation is essential to the credibility of the resulting model.

Software

Two major concerns of the customer were the expense of modelling software and the amount of time necessary for members of the Workforce Assessment Team to become proficient model users. The WAT already maintains a contract for SAS (Statistical Analysis System), so it was initially chosen for all necessary multivariate analysis and logistic regression fitting. For relatively small data sets, however, the very large SAS system is often cumbersome to operate. A second statistical software package called JMP was investigated as an alternative to SAS.

JMP is a software package for interactive statistical graphics which includes a wide variety of data analysis tools. It is a comparatively new product with its first release in October, 1989. Completely developed by the SAS Institute Inc., but not part of the SAS System, JMP offers point-and-click methods of logistic regression fitting and readily accomplishes data normality tests, principal component analysis, and outlier analysis with an easy-to-learn and easy-to-use interface. When working in JMP, the data is contained in spreadsheet for viewing, editing, entering, and manipulating. Since the SAS System, which the customer is already proficient at, is capable of reproducing all of JMP's techniques, and the size of the data tables required for this thesis project are relatively small, JMP is a desirable alternative to the mammoth SAS System.

The end-product desired in this thesis is a user-friendly application which not only allows the Workforce Assessment Team to forecast its future civilian inventory, but also output the desired forecasts as text files, aggregated tables, or even charts. Several of the spreadsheet packages on the market are capable of meeting these requirements. However, the front-end model interface was developed as a Microsoft Excel template because the customer already uses this software.

III. Methodology

Data Collection

The NSA provided me with an initial data set drawn by systematically sampling from their personnel archives. Records from 1989 to 1996 were sorted on social security number and every n^{th} record was selected. For security reasons, the value of n was not divulged. After the records were selected, all action codes that existed for a particular individual were output to a data file. Many of the individuals whose records were selected had several action codes due to multiple personnel actions. Multiple personnel actions weren't uncommon due to the eight-year time span. A selected individual's record may have indicated a promotion in the first year, a migration in the third, and a retirement in the eighth. Because each of these personnel actions happened in different years, I chose to consider each of the different actions as separate records. For example, a single record with action codes identifying a promotion in the first year and a migration in the third was treated as two individual records; one with a promotion and the other with a migration. The data set initially consisted of 4786 records each with 15 fields. After manipulating the records with multiple action codes, the number of records was expanded to 9832.

The 15 fields (shown in Table 2) screened from the NSA databases were modified to create the 13 variables (shown in Table 3). The first five variables, Bsc_Loss, Plcy_Lss, ECF, EGRADE, and New_Hire, are responses for the personnel functions to be modelled. For the binary variables, an action (loss or new hire) was identified by a response value of one while an inaction was identified by a zero value. Each record was assigned values for all response variables. The nominal response variables, ECF and EGRADE, contain the end-of-year career field and grade respectively. The other eight variables will be used as predictors of the responses. All five personnel functions to be modelled will use these variables to assure common dimensionality.

Common dimensionality is necessary because the five regression models will be used to create a single personnel inventory forecasting application. In the application, personnel inventory variables will have common dimensionality to facilitate proper flow of the civilian personnel inventory from one personnel function to another. Each of the major personnel functions will be modelled with logistic regression. Figure 5 depicts the flow of NSA civilians through major personnel functions using these multidimensional variables and the supermatrix concept.

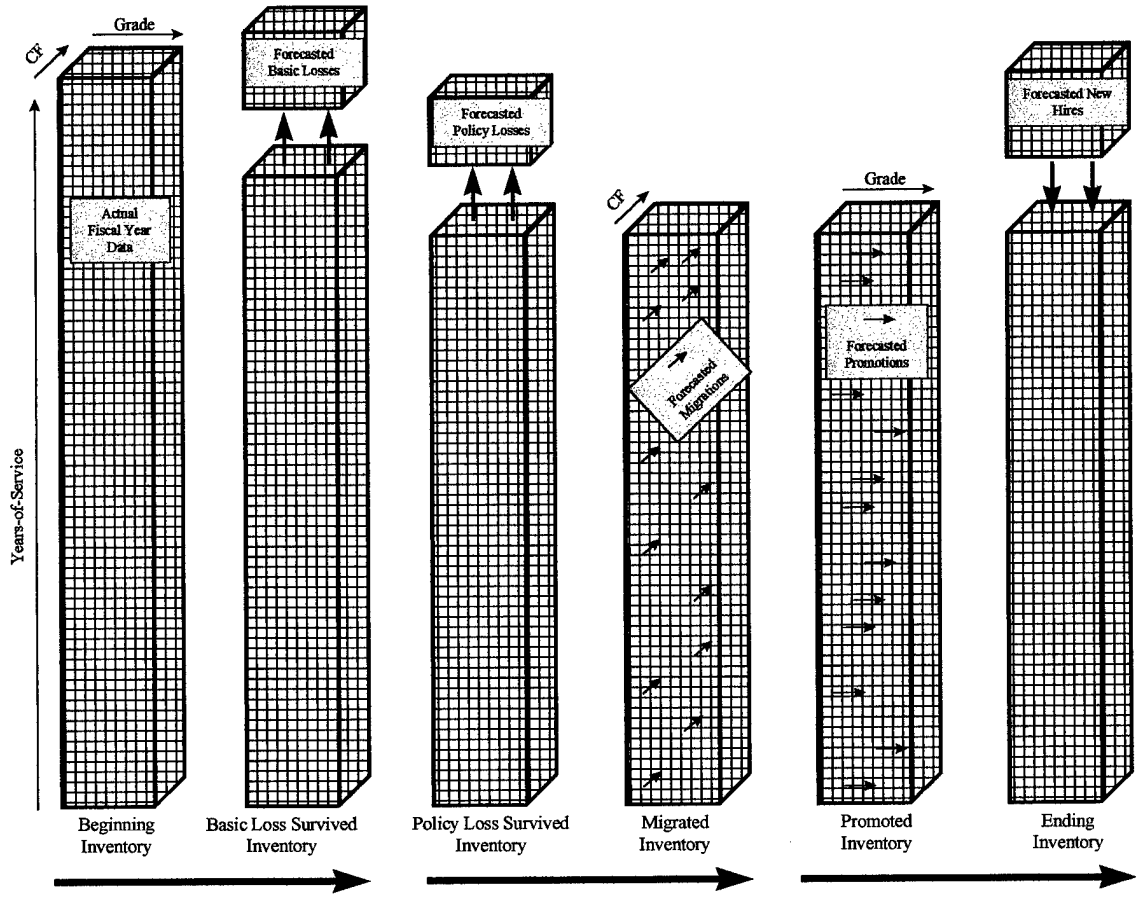


Figure 5. Civilian Personnel Flow

The beginning-of-year career field and grade predictor variables (BCF and BGRADE) were defined with a nominal class similar to their end-of-year counterparts (ECF and EGRADE). The grade variables contain values from 1 through 16 corresponding to an individual's grade. One modification was made regarding Senior Executive Service (SES) grades. Since less than 1% of the NSA civilians hold an SES grade, they have been merged into a single value. All records which indicated an SES grade were given the highest grade of

16. The career field variables have values 0 through 9, corresponding to the first digit of an individual's Career Occupation Specialty Code (COSC). Table 4 shows general identifiers for each of these ten digits.

Table 4. Career Field Identifiers

CAREER FIELD	GENERAL IDENTIFIER
0	Directorate/Management
1	Signal/Language Interpretation
2	Technical
3	Administrative Support
4	Technical/Logistic Support
5	Information Security
6	Office Support
7	Medical Support
8	Intern/Student
9	Maintenance Support

The predictor variables AGE, YOS, and TIG were assigned a continuous class status for obvious reasons. Even though these three variables contain only integer values, they represent continuous personal attributes associated with each record. The GENDER variable is binary where male/female is indicated with a 0/1 respectively. An exception to this rule is where gender is unknown. Since males represent 57.9% of the sample and there were only a handful of records with a code U (unidentified race or sex), I summarily assigned them the male gender. The RACE variable is also binary where white/non-white is indicated with a 0/1 respectively.

The EDCAT variable is nominal where the four categories are indicated according the definitions given in Table 5.

Table 5. Education Categories

EDCAT	CATEGORY
0	Some HS but no degree (and unknown)
1	Some college (certificate or RN) but no degree
2	Degree (Assoc., Bchl's, Mstr's, Law, Profsnl)
3	Degree (Doctoral)

The original data fields LOCUMSAL and RSC contained data with numerous missing values. These two fields were included in the initial data set on the outside chance that they would provide some additional predictive power. Both were removed from further consideration due to the many missing values.

The logistic regression models that will predict each of the major personnel functions require independent predictors. A quick survey of the input variables revealed that not all of the variables are independent of each other. Specifically, AGE, YOS, TIG, and BGRADE are clearly related to each other. The sample data correlation matrix in Table 6 confirms this intuition with coefficients as high as .7449 observed for the combinations of these four variables. Traditionally, correlated predictors would be removed from the set of potential predictors. These four, however, were very predictive of the personnel functions to be modelled.

Principal Component Analysis (PCA) was used to transform these four correlated variables into independent linear combinations for use in the regression models.

Table 6. Correlation Matrix of Predictors

VARIABLE	AGE	YOS	TIG	BGRADE	GENDER	RACE	BCF	EDCAT
AGE	1	0.7449	0.4071	0.5397	-0.204	-0.0273	-0.252	-0.1099
YOS	0.7449	1	0.5047	0.5934	-0.199	-0.0562	-0.3302	-0.0605
TIG	0.4071	0.5047	1	0.1764	-0.128	-0.0907	-0.1445	-0.0231
BGRADE	0.5397	0.5934	0.1764	1	-0.304	-0.206	-0.6263	0.1265
GENDER	-0.204	-0.199	-0.128	-0.304	1	0.0883	0.1749	-0.0006
RACE	-0.027	-0.056	-0.091	-0.206	0.0883	1	0.1883	-0.1043
BCF	-0.252	-0.330	-0.145	-0.626	0.1749	0.1883	1	-0.1308
EDCAT	-0.109	-0.061	-0.023	0.1265	-0.001	-0.1043	-0.1308	1

PCA is normally used to reduce a set of input variables into a smaller, more manageable set of independent linear combinations of the original variables. This reduction of variables comes at the price of lost information contained in the lower order principal components that are ignored. In this case however, the primary concern was not with the number of input variables, but with the independence of the input variables. Therefore, PCA was used to transform the four dependent original variables into four independent principal component scores. Because all four principal components were retained, no information contained in the original variables was lost. Table 7 shows the PCA output.

Table 7. Eigenvalue/Eigenvector Matrices

EIGENVALUE:	2.5273	0.8314	0.4089	0.2325
Percent:	63.1826	20.7841	10.222	5.8113
CumPercent:	63.1826	83.9667	94.1887	100

EIGENVECTORS:	PC1	PC2	PC3	PC4
AGE	0.54873	-0.082	-0.6412	0.53014
YOS	0.57645	0.00046	-0.1679	-0.7997
TIG	0.39022	0.80754	0.39508	0.19879
BGRADE	0.46295	-0.5841	0.63607	0.19982

The eigenvalue indicates the proportion of total variation in the data accounted for by its associated eigenvector. The first eigenvector, for example, accounts for 63.18% of the total variability as indicated by its associated eigenvalue. The third and fourth eigenvectors account for only 10.2% and 5.8% of the total variability respectively. In principal component analysis, these two would be seriously considered for removal as a data reduction technique. In this case, however, PCA is being used only to construct independent variables from the original correlated data.

To understand the reason data reduction was not accomplished at this point, consider that each of the five logistic regression models will be affected differently by AGE, YOS, TIG, and BGRADE. The eigenvector values in Table 7 indicate that each of these original variables are primarily associated with different eigenvectors. For example, the YOS variable with an eigenvector value of

-0.7997, identifies strongly with the fourth principal component. Considering the major personnel functions, however, YOS will certainly be valuable in predicting retirement basic losses. The data reduction will be accomplished during the stepwise regression fitting procedure as these four independent principal components along with the remaining four original variables will serve as potential predictors for each of five the personnel functions to be modelled.

To be certain that the principal component scores and the remaining original variables were independent, correlations between the new principal component scores and the four other predictor variables were calculated. Table 8 confirms that the four principal component scores are uncorrelated with each other (as expected) and that, with only a few possible exceptions, no other practical correlations exist between any of the eight predictor variables.

Table 8. Correlation Matrix of Modified Predictors

VARIABLE	PC1	PC2	PC3	PC4	GENDER	RACE	BCF	EDCAT
PC1	1	0	0	0	-0.2623	-0.112	-0.4246	-0.0287
PC2	0	1	0	0	0.0993	0.0541	0.2957	-0.0916
PC3	0	0	1	0	-0.1246	-0.2188	-0.3728	0.2376
PC4	0	0	0	1	-0.0724	-0.0596	-0.0486	0.0224
GENDER	-0.2623	0.0993	-0.1246	-0.0724	1	0.0883	0.1749	-0.0006
RACE	-0.112	0.0541	-0.2188	-0.0596	0.0883	1	0.1883	-0.1043
BCF	-0.4246	0.2957	-0.3728	-0.0486	0.1749	0.1883	1	-0.1308
EDCAT	-0.0287	-0.0916	0.2376	0.0224	-0.0006	-0.1043	-0.1308	1

Several of the variable pairs from Table 8 could be considered statistically dependent based upon correlation coefficients with magnitude greater than 0.20. Closer investigation, however, indicates that these statistically significant correlations are not practical in light of the variable definitions. The BCF (beginning-of-year career field) variable, with a correlation coefficient of -0.4246, appears to be correlated with the first principal component. This initial evidence is misleading because BCF is a nominal variable with arbitrarily assigned categories, and the first principal component is a linear combination of AGE, TIG, YOS, and BGRADE. For BCF, the association of the ten actual career fields with variable categories (0-9) is completely arbitrary and does not imply ordinality. Any apparent statistical correlations between BCF, or any of the nominal variables, and the four principal component scores could be eliminated by carefully renumbering the categories. The actual personal attributes represented by the nominal variables are not practically correlated with the principal component scores so they are taken to be appropriate for use in regression modelling.

A lengthy investigation of the data led to many modifications to the records. A total of 32 records were deleted because they contained missing or incorrect data. Some records with minor inaccuracies or obvious omissions

were corrected. This initial investigation was driven by simply looking at the data in tables and histograms. After modifying or deleting the incorrect records, an outlier analysis using Mahalanobis distance was accomplished.

A quick look at the frequencies of actions indicated that hires were grossly over-represented. There were 550 basic losses, 400 policy losses, and 945 new hires over the same eight year period. The Workforce Assessment Team confirmed my hypothesis that basic losses and new hires normally offset each other, and that twice as many hires were inadvertently selected. To remedy the problem all hires were assigned a uniform (0,1) random number, sorted on that random number, and then broken into two groups of 472 and 473. The first group was left in the data set and the other saved for validation. The remaining data set contained 9327 records.

Since the NSA had requested forecasts of full-time employees only, the data set was also stripped of all records that didn't reflect a full-time category of employment (EMPCAT). After sorting on the EMPCAT data field, all records with the following values were removed:

I -- intermittent
L -- full time on extended leave w/o pay
P -- part time
R -- part time on extended leave w/o pay.

Removal of these 231 records left the sample size at 9096.

All records which had no actions over the eight year time period lacked an action code and action date. These records were given an action date of 09/30/96 which was the last day of the time period. An action code of 000A, indicating no action, was also assigned. This combination of action date and action code reflected the fact that for the entire period no personnel activity occurred. If a record, which was not a migration, was missing an end-of-year COSC then the beginning-of-year COSC was inserted. If the date-of-birth indicated a (non-leap year) 29 Feb birthday then it was changed to 28 Feb. If the end-of-year grade was missing on a record which was not a promotion, the beginning-of-year grade was inserted. After making many corrections similar to the ones mentioned above there remained 35 records which did not lend themselves to obvious interpretation. These records were deleted leaving the remaining data set with 9061 records.

To be sure that the accuracy of the data set had been thoroughly validated, additional tests were performed. The multivariate nature of the records necessitated a more technically proficient method of investigation than visual inspection. To complete the record investigation, an outlier analysis was accomplished by plotting the Mahalanobis distance between each point and the multivariate

mean (centroid). Figure 6 contains a plot of all 9061 records by their respective Mahalanobis distance.

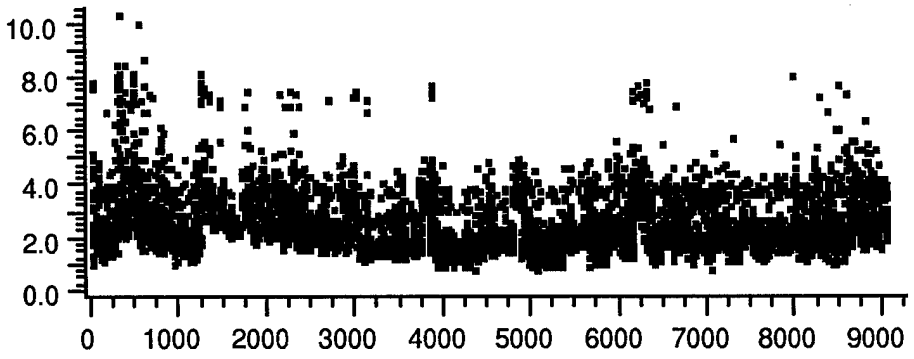


Figure 6. Plot of Records by Mahalanobis Distance

This plot clearly identifies two distinct strata in the 9061 records. The points in the upper strata, all with Mahalanobis distances of six or greater, appear to be highly-suspect data points. However, upon closer inspection, these 123 records correspond to two groups of individuals with uncommon, but valid, predictor variable values. Eighty of the points in the upper strata correspond to individuals with an education category of zero (no high school diploma). Since those records represent a very small portion of the population, their values didn't have a significant effect on the location of the centroid. Hence, when these records' unusual values were compared to the multivariate mean, they were quite distant from the mean and appeared to be invalid. Despite the large Mahalanobis

distances of these points, the data is found to be an important link to a relatively small group of records, without which a portion of the model's capability would be lost.

An important note is that all database records with EDCAT values of zero are contained in this group of outliers, thus confirming the belief that this group contains data that must be included in subsequent models. The other 43 records that also fell into this outlier group were all significantly older than the average civilian employee. This second set of outliers had a mean age of 64.8 and a mean time-in-service of 38.3 years. When compared to the population age and time-in-service averages of 41.4 and 16.5 respectively, the records appear to be a valid minority. In all, many records were investigated as part of the outlier analysis, but all of the points proved to be valid and appropriate for use in model fitting. For further investigation of the outlier points see Appendix B for a list of data records with Mahalanobis distances greater than 6.0.

With independent variables and accurate database records having been established, the next step was to fit logistic regression models to the data. Since each of the personnel functions to be modelled had its own response variable, five separate models were fit. A stepwise

procedure was followed in each case to determine the subset of predictor variables which produced the most predictive models.

Regression Fitting

The set of modified records and variables was saved as a JMP database from the spreadsheet window. The database of 9061 records had the five response and eight predictor variables discussed previously. The binary response variables were assigned values of 0 or 1 (for example, Y=1 if a basic loss occurred; otherwise Y=0). The linear logistic model for binary response variables has the form

$$logit(p) = \alpha + \beta' x$$

where x is a vector of explanatory variables, $p = \Pr(Y=1|x)$, α is the intercept parameter, and β is the vector of slope parameters. Using the estimated parameters, p can be calculated as follows:

$$p = \frac{e^{logit(p)}}{(1 + e^{logit(p)})} .$$

When the response variable is nominal with r categories then $p_r = \Pr(Y=j|x)$ for $j = 1,\ldots,r-1$. Consider the ECF response variable which has $r = 10$ categories. Logit scores will be calculated for p_0 through p_8 where the control group is p_9. Using the estimated parameters, the nine logit scores $(p_1,\ldots,p_8)$ can be calculated as before. Because category 9

is the control group it follows that $logit(p_9) = 0$ and $p_9 = 0.50$. Resultantly, if p_1 through p_8 are all less than 0.50, then the control group category is the most likely end-of-year career field. Otherwise, the category associated with the greatest p_j is the most likely end-of-year career field.

JMP didn't offer stepwise regression for logistic regression so I accomplished it manually. In each of the five cases, the response was initially fitted against the full set of predictors. The least significant of the predictors, with a p-value greater than 0.10, was iteratively removed until all those which remained were significantly different from zero. In regression analysis the p-value associated with a predictor indicates the probability that the variable's effect on the response is no different than zero. Selecting predictors with small p-values prevents including nonpredictive or "noise" variables. The procedure is stopped when all of the remaining predictors have p-values less than some predetermined value - in this case 0.10.

One subtlety of the stepwise procedure is that the successive p-values obtained in each iteration are highly correlated. Another problem with this procedure is that no multicollinearity is assumed to exist (the predictors are assumed to be absolutely independent). Even though I paid considerable attention to investigating and correcting

multicollinearity in this data set, some small interdependencies certainly remain. With these procedure and data imperfections in mind, I chose to include a few variables with p-values greater than 0.10 in final models despite their high test scores. These exceptions occurred where the p-values were close to 0.15 for variables which I felt could actually be predictive. These case-by-case decisions were made to utilize variables which may have been under-predicting due to sample variation. The JMP output for the final fitted models are contained in Appendices C through G.

The four correlated variables used in the PCA loaded differently on each of the four principal components. As a result, the fitting procedure for each of the regression models was capable of selecting a different subset of the four principal component scores as the best predictors. For example, the principal component score which the TIG variable loaded most heavily on was considerably more predictive of promotions than migrations. The new hires model was the one personnel function which did not use the principal component scores as potential predictors. Since AGE and EGRADE are not strongly correlated, the original variables were used for predictors.

Labels were given to the four principal component scores to aid in their interpretation. Based upon the

loadings of the original variables on the four principal components the labels were assigned as follows:

PC1	Maturity Index
PC2	Experience & Grade Contrast
PC3	Grade & Age Contrast
PC4	Age & Seniority Contrast.

Table 9 summarizes the regression fitting results by identifying the significant variables included in each of the final logistic regression models.

Table 9. Significant Predictors of Personnel Functions

VARIABLE	BASIC LOSS	POLICY LOSS	MIGRATE	PROMOTE	NEW HIRE
Maturity Index	X	X	X	X	
Experience & Grade Contrast	X	X	X	X	
Grade & Age Contrast	X		X	X	
Age & Seniority Contrast	X		X	X	
BCF	X	X	X		
GENDER	X		X	X	X
RACE	X			X	X
EDCAT	X				X
ECF					X
EGRADE					X

Initial Validation

To verify the predictability of each of the fitted component logistic regression models, a randomly selected 1000 record holdout set was used for validation. This initial validation was performed according to methods discussed by Law and Kelton (1991: Ch 5). After fitting each of the models with the reduced data set, the estimated parameters were applied to the 1000 held-out records. The known responses were then compares to those predicted by the model. The binary logistic regression models for basic losses, policy losses, and new hires produced the results shown in Figure 7.

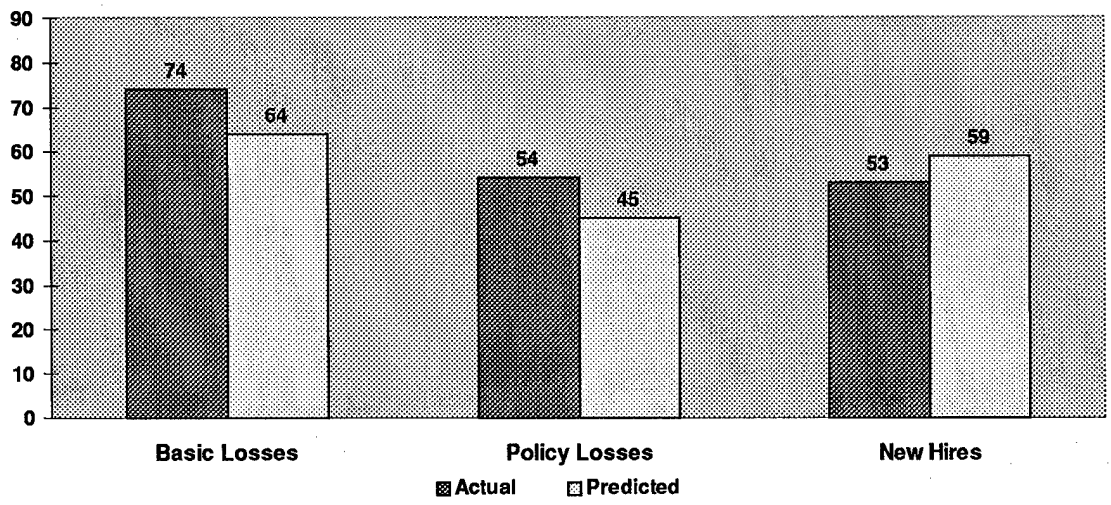


Figure 7. Validation Results for Binary Response Models

The nominal logistic regression models for migrations and promotions produced the results shown in Figures 8 and 9 respectively.

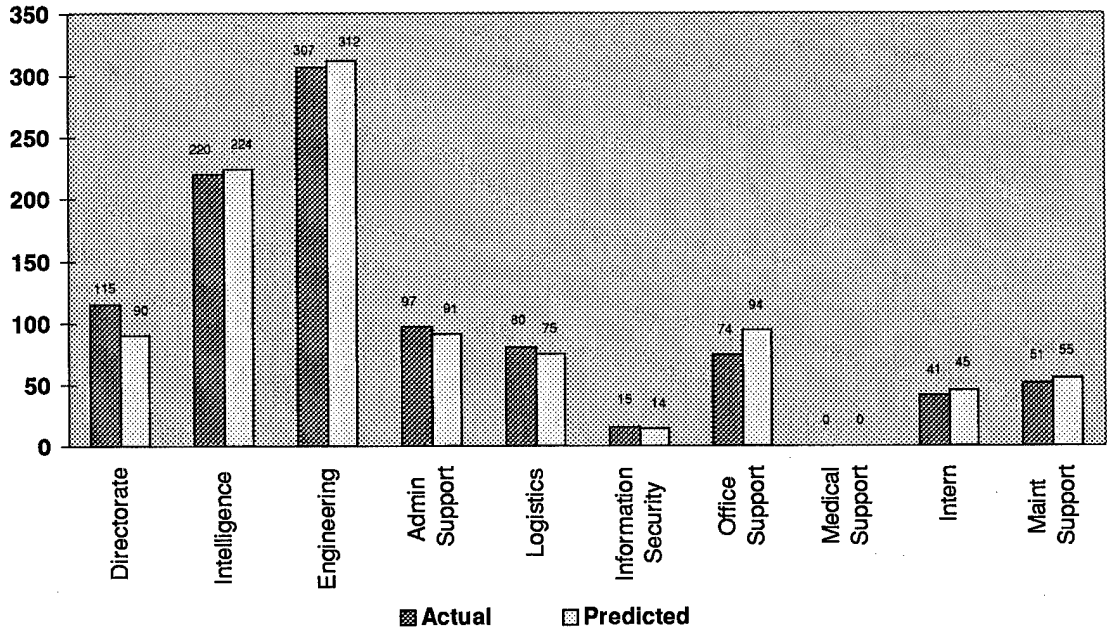


Figure 8. Validation Results for Migrations Model

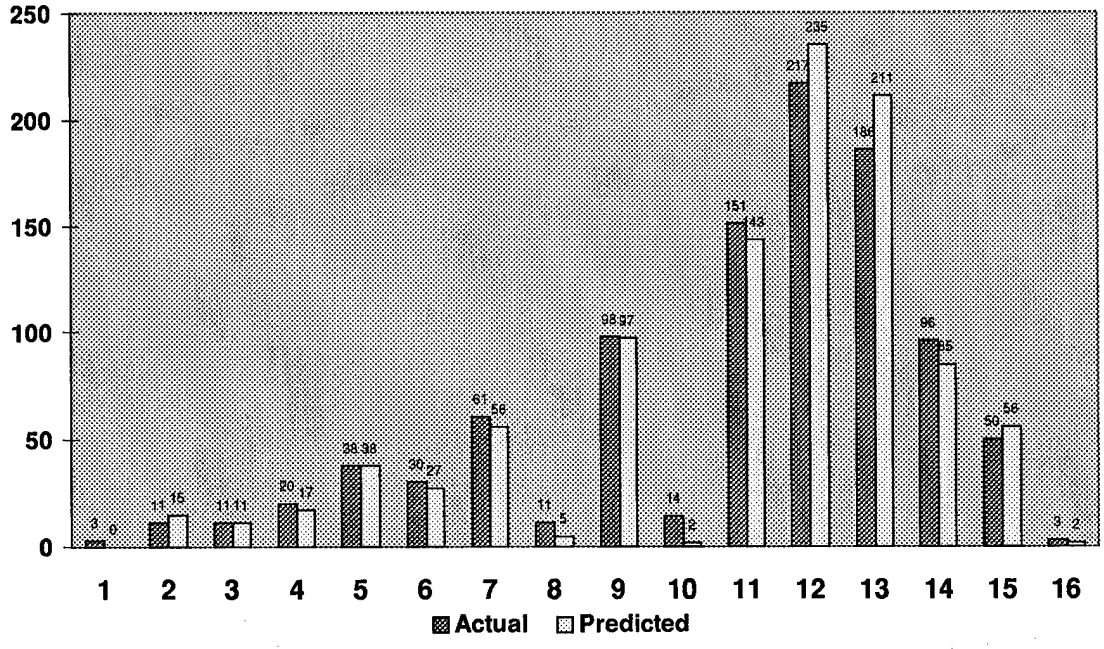


Figure 9. Validation Results for Promotions Model

Results of this initial validation indicate that the individual logistic regression models are capable of predicting their respective responses with only small to moderate error margins. Based on these results, the parameters for the five regression models were re-estimated using the entire data set for use in the spreadsheet tool. These parameter estimates will drive the calculations in the Microsoft Excel template.

Application Building

A Microsoft Excel template was developed to utilize the logistic regression parameters calculated in JMP. The template contains several worksheets which clearly display the modelling stages shown previously in Figure 1. Formulas were entered in the appropriate cells and were then verified against the 1000 record holdout set. This was accomplished by comparing the regression validation data from JMP with the Excel model calculations (using the same 1000 records) to verify that the model was computing the correct values. Figure 10 shows the initial display of the Excel template.

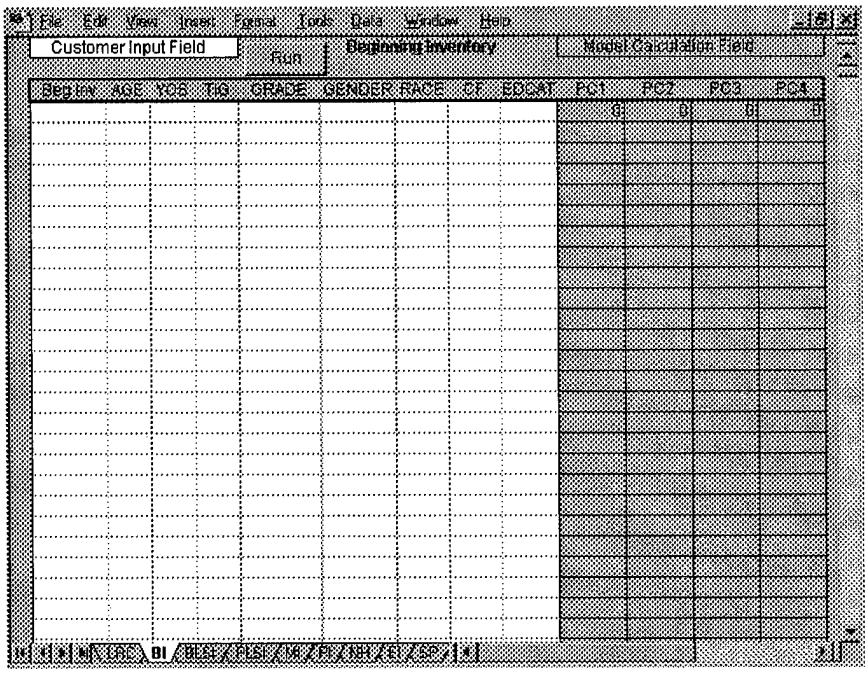
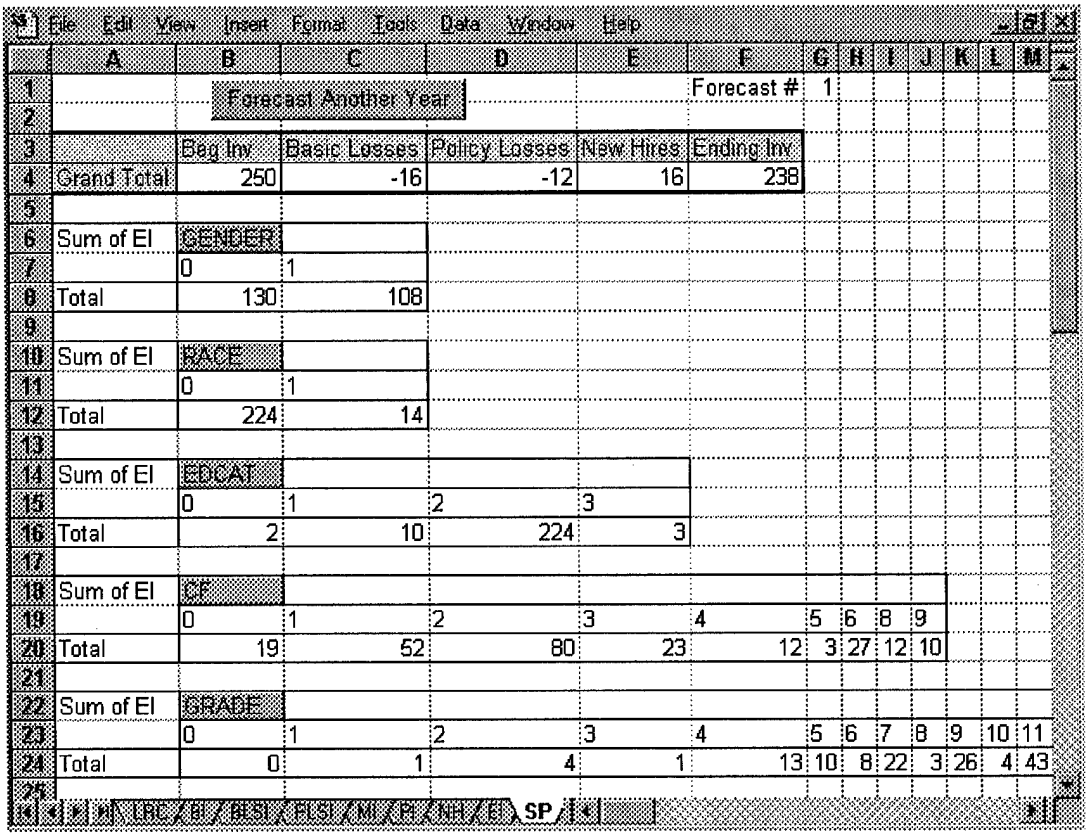


Figure 10. Microsoft Excel Template

The BI (Beginning Inventory) worksheet accepts the customer's only required input. The beginning inventory data must be collected from the NSA's most recent 30 September's EDPSMSTR file. The EDPSMSTR file contains a snapshot of the NSA's end-of-fiscal year civilian personnel inventory. Some minor manipulations will be necessary before the raw EDPSMSTR data extract can be entered into the BI worksheet. Apart from formatting the text file, there are no other mandatory requirements. It is recommended, however, that the customer spend some time investigating the validity of the data set. Records with missing or erroneous entries should be corrected or deleted. Once the customer

has entered a valid beginning inventory, a single-year forecast can be accomplished by selecting the "Run" button.

The LRC (Logistic Regression Coefficients) worksheet contains all of the estimated parameters. The coefficients should be updated annually as data for additional fiscal years becomes available. Updating the coefficients requires that the customer refit the JMP data set (including the new fiscal year data) and then simply copy them into the appropriate cells in the LRC worksheet. A macro called "PullDown" is executed upon selection of the "Run" button. This macro copies the formulas necessary for each worksheet to the appropriate cells. The "PullDown" macro also calculates aggregate worksheet totals on the SP (Scratch Pad) worksheet. A copy of the Visual Basic code for all of the spreadsheet model macros are contained in Appendix H. Figure 11 shows a portion of the standard output calculated on the SP worksheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Forecast Another Year				Forecast #	1						
2													
3		Beg Inv	Basic Losses	Policy Losses	New Hires	Ending Inv							
4	Grand Total	250	-16	-12	16	238							
5													
6	Sum of EI	GENDER											
7		0	1										
8	Total	130	108										
9													
10	Sum of EI	RACE											
11		0	1										
12	Total	224	14										
13													
14	Sum of EI	EDCAT											
15		0	1	2	3								
16	Total	2	10	224	3								
17													
18	Sum of EI	CF											
19		0	1	2	3	4	5	6	8	9			
20	Total	19	52	80	23	12	3	27	12	10			
21													
22	Sum of EI	GRADE											
23		0	1	2	3	4	5	6	7	8	9	10	11
24	Total	0	1	4	1	13	10	8	22	3	26	4	43
25													

Figure 11. Standard Output on Scratch Pad

The BLSI (Basic Loss Survived Inventory) and PLSI (Policy Loss Survived Inventory) worksheets do not require input. If the customer wishes to affect the prediction of basic or policy losses, however, a "Modify Losses" button is available. Selecting one of these buttons executes a succession of dialog boxes that prompt the user for input and then make notification of the resulting change. The MI (Migrated Inventory) and PI (Promoted Inventory) worksheets neither require customer input nor offer control lever modification. The logit scores which determine the new

career fields and grades are calculated from the logistic regression coefficients and the variable values of each record. The highest logit score for a particular record, corresponds to the nominal response category which that record is most likely to be in. The MI and PI worksheets affect the internal distribution of the single-year forecast, but not the size. Because the basic and policy losses are both partially predicted by career field and grade, the accuracy of a multiple-year forecast is dependent upon the migrations and promotions accomplished in these two worksheets.

New hires calculated on the NH worksheet are added to the promoted inventory on the EI (Ending Inventory) worksheet. The AGE and YOS variables are incremented on the EI worksheet to simulate the passage of time. The aggregate level of new hires can be adjusted by selecting the "Modify New Hires" button on the NH worksheet. The customer can change the new hires control lever in the same manner as for basic and policy losses.

To accomplish a two-year forecast the customer must copy the single-year forecasted ending inventory into the BI worksheet and then select the "Run" button. This process may iterate until the desired number of years of the NSA's civilian personnel inventory are forecasted, hence the name Iterative Civilian Forecast Model. The customer may

automatically accomplish these steps by selecting the "Forecast Another Year" button. This button, located on the SP worksheet, is hooked to a macro called "ForecastAnother" that executes all necessary commands.

IV. Model Adequacy

The initial validation accomplished previously used a holdout set of 1000 records. The purpose was to make sure that each of the five logistic regression models were acceptable. In this section a completely distinct set of records were utilized to validate the spreadsheet model. There is a subtle difference between these two attempts at validation. As was shown previously, the initial validation indicated that each of the regression components were predictive. However, that result alone doesn't validate the spreadsheet model. To determine the adequacy of the spreadsheet model, which incorporates customer input and the five regression models' coefficients, a second validation was accomplished.

At my request, the Workforce Assessment Team collected additional personnel data for this secondary validation. The method of collection was similar to that of the first data set with one modification. Records from the end-of-year snapshots for FY95 through FY97 were collected in equal proportions and used to create three text files. The size of each year's complete snapshot was not provided, but each of the data sets were large enough to be considered representative samples. The data files were checked for incorrect data field entries and outliers before variables

were created. Minor record modifications were accomplished in a manner similar to that discussed previously (for the initial data set).

One-Year Validation

Since the sample data sets represented three consecutive fiscal years in proportion , I used them to produce one and two-year validation comparisons. The FY95 data, which had 3166 records, was utilized in the ICFM as a beginning inventory. A single-year forecast was made and the output compared to the FY96 data. The metric used for determining model accuracy is percent misclassification. Because the model does not track individuals by social security number, misclassification must be calculated at an aggregated level. For example, if the model incorrectly identifies five new hires with high school diplomas as college graduates, then these two categories of the EDCAT variable will be off by five. Because the model does not track individuals, however, my method of calculating misclassifications is considerably more conservative. I simple aggregate the absolute error between the actual and predicted at each category. For the given example, I would have counted the five overage errors in the college category as well as the five underage errors in the high school

category, resulting in a doubled misclassification percentage.

The model's inability to track individuals is due to the NSA's unwillingness to divulge social security number as part of the original data. Calculating the accuracy metric in this conservative fashion, however pessimistic it may be, still provides good insight into the models predictive ability. The one-year validation comparisons are shown in Table 10.

Table 10. One-Year Validation Percent Misclassification

ONE-YEAR VALIDATION	AGGREGATE INVENTORY	EDCAT	GENDER	RACE	CAREER FIELD	GRADE	YOS	AGE	TIG
PERCENT MISCLASSIFICATION	0.64	1.06	1.75	2.51	5.73	7.28	14.96	15.38	18.63

A misclassification percentage of .64% for the aggregate inventory initially indicated the model was working well. Similar comparisons were accomplished at each level of model resolution. All of these one-year validation comparisons indicated that the model predicted quite accurately. Comparisons at higher resolution than the aggregate inventory yielded less accurate, yet still acceptable, predictions. Validating the ICFM's multiple-year forecasting ability required additional data and model runs.

Two-Year Validation

The single-year forecast which resulted from the one-year validation was used as the beginning inventory for a second run of the ICFM. The ending inventory for this second iteration represented a two-year forecast from the original starting point. The initial input was from FY95, so the two-year validation used output from the second run for comparison against the FY97 data set. Table 11 displays these two-year validation comparison.

Table 11. Two-Year Validation Percent Misclassification

TWO-YEAR VALIDATION	AGGREGATE INVENTORY	EDCAT	GENDER	RACE	CAREER FIELD	GRADE	YOS	AGE	TIG
PERCENT MISCLASSIFICATION	2.51	2.51	3.75	2.51	7.06	9.23	15.89	20.13	29.39

A misclassification percentage of 2.51% for the aggregate inventory means that the model predicted a force of 2884 when the actual value was 2958. The remaining two-year validation comparisons reinforce the models credibility. As was the case for the one-year forecast, comparisons at higher resolutions yielded less accurate predictions. Figure 12 compares the percent misclassification at the various levels of model fidelity.

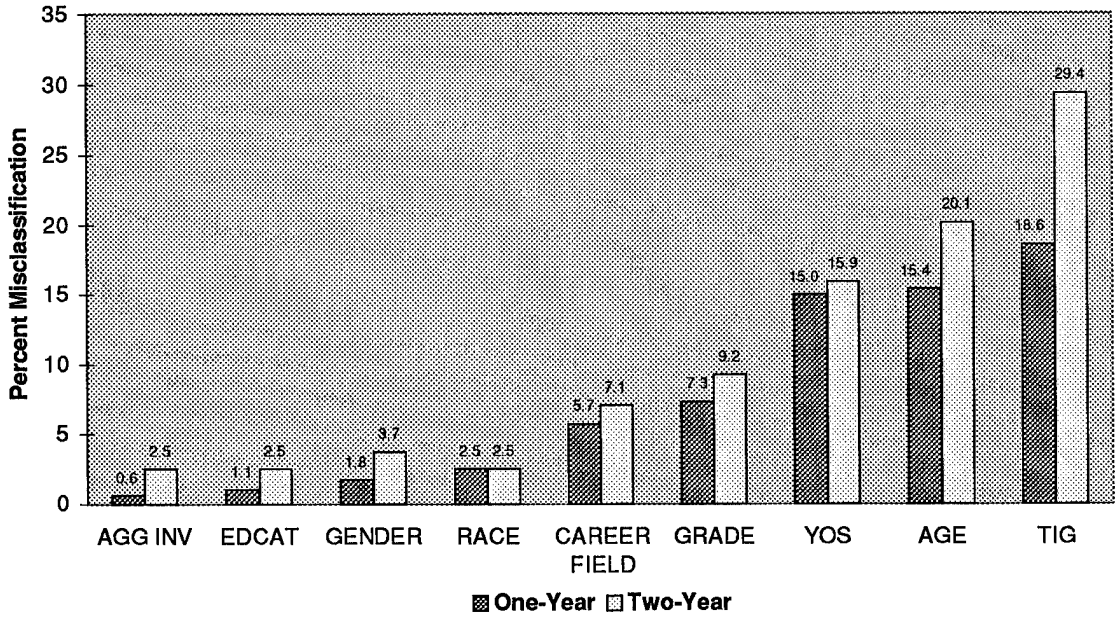


Figure 12. Comparison of One and Two-Year Validation Results

As was expected the two-year predictions were less accurate than the one-year predictions. Another trend identified by Figure 12 is that an increased degree of resolution was usually characterized by a higher percent misclassification. Increased misclassification rates not withstanding, the two-year forecast was still quite good. So based upon the preceding analysis, the ICFM has been shown capable of producing multiple-year forecasts of the NSA's civilian personnel inventory in the aggregate as well as for various levels of higher fidelity.

Run Time

While producing the necessary forecasted inventories for the validation process the average run-time (for the

3166 records data set) was less than 3 minutes. The second year of the two-year forecast usually took just under 19 minutes. I believe the increase in processing time for the second year was due to memory congestion. The computer on which the validation runs were completed had a Pentium 120 MHz processor and 16 MB of memory. An additional test-run was made with a considerably larger data set to verify the model's ability to function with a realistic number of records. Run-times for this test data set, which had 12,500 records, were approximately 22 and 89 minutes for the first and second years respectively. While these times are much higher than those of the validation runs I believe that they are still viable.

V. Conclusions

Recommendations

While the model developed in this thesis has been shown to be valid for its intended application, there still remains room for improvement. Since the data used to fit the logistic regression parameters was drawn proportionately from the population, the resulting spreadsheet predictions are subject to sample variability. Security issues, which necessitated the data sampling, do not affect the NSA's Workforce Assessment Team. While not essential, refitting the regression parameters using the entire data set would certainly be worthwhile. Changing personnel policy and evolving internal dynamics will erode the model's predictability over time. So even if the regression parameters (using the entire data set) are not immediately recalculated by the WAT, this task should be accomplished periodically to prevent the model from becoming obsolete.

The spreadsheet model is compatible with both Windows®95 and Windows®97 versions of Excel, but future releases may require that some minor changes be made for proper operation. Since these new versions usually offer the user more options as well as increased efficiency, it is recommended that the NSA investigate upgraded software as it becomes available. The Visual Basic subroutines are easily

modified to accommodate either changes in the personnel process or software upgrade. A particularly simple modification to the spreadsheet model would be the addition of forecasted payroll estimates. Since the model already forecasts civilian personnel inventory by grade, producing payroll estimates would require only a pay chart and some simple arithmetic.

Perhaps the best mechanism for improvement is an annual validation of the model. Analyzing the model's accuracy trend on a yearly basis would ensure that minor discrepancies receive attention before becoming unmanageable. Accomplishing any minor maintenance identified by the validation analysis would improve accuracy and possibly run-time efficiency. Another benefit of an annual validation process would be the credibility gained by decision makers as the model was shown to provide consistently accurate forecasts.

Summary

The objective of this thesis was to utilize both historical data as well as the NSA's human resource expertise and develop a model to accurately forecast its civilian personnel inventory for multiple years. The historical data provided by the NSA was scrutinized to identify and correct any invalid data field entries. After

the data entries were verified or corrected, an investigation of outlier records was accomplished. The resulting database was used to build five logistic regression models. These five fitted regression models (basic losses, policy losses, migration, promotions, and new hires) represented the major personnel functions faced by the NSA's civilian workforce. Outlier analysis, regression fitting, and initial validation was accomplished using the JMP statistical software package.

While initial validation showed that the regression components were accurate, no method existed which allowed expert opinion to be utilized. Incorporating the customer's expertise into the forecasting process required that these five regression model components be merged into a single model. Excel, Microsoft's spreadsheet software, was used to satisfy these two requirements resulting in a single model that accepts customer control lever settings. With these control levers the customer can affect the aggregate number of basic losses, policy losses, and new hires for any forecast year. A secondary validation of this spreadsheet model showed it to be valid for multiple-year forecasts.

The forecasted civilian personnel inventories can be output in the aggregate or in any combination of the following eight model dimensions: age, years-of-service, grade, time-in-grade, gender, race, career field, and

category-of-education. Charting capabilities built into Excel make specialized report building quite simple. The combination of the model's validated accuracy and Excel's user-interface provides a model that meets all of the original requirements which I sincerely hope the NSA's Workforce Assessment Team will use to help them do its job easier and better.

Appendix A: Field Names of NSA Databases

FIELD		DATABASE		
NAME	DEFINITION	EDPSMSTR	ACTIVITY	PAST EDPS
ACD	AGENCY COMPUTATION DATE	X		X
ACDF	AWARD COMPUTATION DATE	X		X
ACT_CODE	ACTION CODE		X	
ACT_DTE	EFFECTIVE DATE OF ACT_CODE		X	
AGE	AGE	X		X
APPTYPE	TYPE APPOINTMENT	X		X
ARD	AGENCY RELEASE DATE	X		X
ASGORG	ASSIGNED ORGANIZATION	X	X	X
ASGORGR	ASGORG RECODE	X	X	X
BILLETORG	BILLET ORGANIZATION	X		
COSC	CAREER OCCUPATION SPEC CODE	X	X	X
COSCLVL	BILLET COSC ASSIGNMENT DESIG.	X		X
COSCR	COSC RECODE	X	X	X
CPERCENT	TSP PERCENTAGE FOR C FUNDS		X	
DETCOSC	DETAIL COSC	X		
DETDTE	DETAIL DATE	X		
DETFUNCTL	DETAIL FUNCTIONAL NUMBER	X		
DETJOBNR	DETAIL JOB NUMBER	X		
DETJOBTLE	DETAIL JOB TITLE	X		
DETORG	DETAIL ORGANIZATION	X		
DETSKILL1	DETAIL LANGUAGE SKILL	X	X	
DETSKILLR	SKILL LEVEL W/LANGUAGE	X		
DLE	DATE OF LAST EQUIVALENCY	X		
DOB	DATE OF BIRTH	X	X	X
DOBR	DOB RECODE	X		
DOG	DATE OF GRADE	X	X	X
DTYORG	DUTY ORGANIZATION	X		X
DTYORGR	DTYORG RECODE	X		X
DTYSPEC	DUTY SPECIALTY (MILITARY)	X		
DTYSPECR	DTYSPEC RECODE	X		
EDLVL	EDUCATION LEVEL	X	X	X
EMPCAT	EMPLOYMENT CATEGORY	X	X	X
EOD	ENTER OF DUTY DATE	X	X	X
F_ASGORG	FROM ASSIGNED ORGANIZATION		X	
F_ASGORG_R	FROM ASGORG RECODE		X	

F_COSC	FROM CAREER OCCUPATION SPEC CODE		X	
F_COSC_R	FROM COSC RECODE		X	
F_CPERCENT	FROM TSP PERCENTAGE FOR C FUNDS		X	
F_DETSKILL1	FROM DETAIL LANGUAGE SKILL		X	
F_DOG	FROM DATE OF GRADE		X	
F_EMPCAT	FROM EMPLOYMENT CATEGORY		X	
F_FPERCENT	FROM TSP PERCENTAGE FOR F FUNDS		X	
F_FUNCTL	FROM FUNCTIONAL CODE		X	
F_GPERCENT	FROM TSP PERCENTAGE FOR G FUNDS		X	
F_JOBNR_6	6TH POSITION OF F_JOBNR		X	
F_JOBNR_89	8TH AND 9TH POSITION OF F_JOBNR		X	
F_LOCADJ	FROM LOCALITY PAY ADJUSTMENT		X	
F_LOCCUMSAL	FROM LOCALITY ACCUM SALARY		X	
F_LOCPCT	FROM LOCALITY PERCENT		X	
F_NTE	FROM NOT TO EXCEED DATE		X	
F_PART_OR_F	PRTTME/FLLTME CODE AFTER ACTION		X	
FIELDS_CHAN	INDICATES FLDS CHANGED BY ACTION		X	
FLSA	FAIR LABOR STANDARDS ACT CODE	X		X
FPERCENT	TSP PERCENTAGE FOR F FUNDS		X	
FUNCTL	FUNCTIONAL CODE	X	X	X
GRD	GRADE	X	X	X
HOURS	HOURS	X	X	
IDENT	IDENT	X	X	X
JOBNR	JOB NUMBER	X	X	X
JOBNR4	4TH POSITION OF F_JOBNR	X	X	
JOBNR56	5TH AND 6TH POSITION OF F_JOBNR	X	X	
JOBNR7	7TH POSITION OF F_JOBNR	X	X	
JOBNR8	8TH POSITION OF F_JOBNR	X	X	
JOBNR9	9TH POSITION OF F_JOBNR	X	X	
JOBTLE	JOB TITLE			X
LBANK	LEAVE BANK STATUS	X	X	X
LOC	LOCATION CODE	X	X	X
LOCDTY	DUTY LOCATION	X		X
MGD	ETHNICITY CODE	X		X
MS	MARITAL STATUS	X		
NAME	NAME	X		X
NAMER	FIRST 5 LETTERS OF LAST NAME	X		X
NATURE_ACT	PRIMARY ACTION EFFECTED		X	

NTE_	NOT TO EXCEED DATE AFTER ACTION		X	
NTEAPPT	EXPIRATION DATE OF NTE APPT	X		X
OPM_DTE	DATE OPM SENT TO OPR		X	
OSROTDT	OVERSEAS ROTATION DATE	X		X
OWE	OFFICER, WARRANT, ENLISTED (MIL)	X		X
PART_OR_FUL	PART-TIME OF FULL-TIME		X	
POSN	THREE DIGIT BILLET NUMBER	X	X	X
PRIMSPEC	PRIMARY SPECIALTY CODE (MIL)	X		X
PRIMSPECR	PRIMSPEC RECODE (MIL)	X		X
PRCESS_DATE	DATE ACTION PROCESSED BY SYSTEM		X	
PS	PAY SCHEDULE	X	X	X
PSRC	PAY SCHEDULE RECODE	X	X	X
PTC	PAY TABLE CODE	X		X
PTCR	PTC RECODE	X		
RANK	MILITARY RANK	X		X
REASON	REASON FOR SEPARATION		X	
ROEORG	RESPONSIBLE OPS. ELEMENT	X		
ROEORGR	ROEORG RECODE	X		
RSC	RECRUITMENT SOURCE	X	X	X
RSC2	RECRUITMENT SOURCE-SECOND	X	X	X
RSC3	RECRUITMENT SOURCE-THIRD	X		X
SALARY	ANNUAL SALARY	X	X	X
SCDF	SERVICE COMP. DATE (FEDERAL)	X	X	X
SCDFL	SERVICE COMP. DATE (LEAVE)	X	X	X
SCDFR	SERVICE COMP. DATE (RETIRE)	X	X	X
SCELVL	SCE/SLE/SLP PAYLEVEL	X	X	X
SCESALARY	ANNUAL SALARY FOR SENIORS	X	X	X
SE	SUBELEMENT	X	X	X
SID	STANDARD IDENTIFICATION	X		X
SKILL1	LANGUAGE SKILL LEVEL W/DIAGRAPH	X	X	X
SKILL1R	SKILL1 RECODE	X		
SPECID	SPECIAL IDENTIFICATION	X	X	X
SRC	SEX RACE CODE		X	
SSN	SOCIAL SECURITY NUMBER	X	X	X
STEP	PAY GRADE STEP AFTER ACTION		X	
STRENGTH	STRENGTH ACCOUNTABILITY	X	X	X
SVC	SERVICE (MIL)	X		X
SVCEXPDT	SERVICE EXPIRATION DATE	X		

TAMT	THRIFT SAVINGS AMOUNT		X	
TENGRP	TENURE GROUP	X		X
TENURE	TENURE GROUP CODE		X	
TITLE	TITLE	X		X
TSPELIG	TSP ELIGIBLE		X	
TSPSTAT	FROM THRIFT SAVINGS PLAN STATUS		X	
VP	VETERANS PREFERENCE	X	X	
WGI	WITHIN GRADE INCREASE	X		X

Appendix B: Mahalanobis Distance Plot Outliers

AGE	YOS	TIG	BGRADE	GENDER	RACE	BCE	EDCAT	M-DIST
60	8	0	1	1	1	9	0	7.96682
56	9	0	2	0	1	9	0	7.76501
64	30	0	6	0	1	9	0	7.84921
59	33	27	13	0	0	2	2	6.78666
56	32	18	14	0	1	2	3	6.43358
59	10	3	9	0	0	4	0	7.26191
68	28	6	1	1	1	9	0	7.87182
66	27	6	2	1	0	9	0	7.6878
68	43	5	2	0	1	9	0	8.54831
68	11	4	2	0	1	9	0	8.13987
68	30	5	2	1	1	9	0	7.8191
71	29	3	2	1	1	9	0	8.09212
72	27	6	2	1	1	9	0	7.90632
66	26	4	2	1	1	9	0	7.72892
38	17	6	7	1	0	9	0	7.14488
70	48	6	7	0	1	9	2	6.28677
67	27	6	9	1	0	9	0	7.37192
73	46	6	10	0	1	9	0	7.91178
60	42	26	11	0	0	2	2	6.20413
75	46	39	11	0	0	4	2	10.4326
71	52	32	11	1	0	1	2	8.25349
60	42	8	11	0	1	4	0	7.3916
66	40	4	11	0	1	9	0	7.82001
60	39	15	12	0	0	2	0	7.2002
66	44	26	12	0	0	1	2	6.15132
59	40	28	12	0	0	1	2	6.85219
62	44	27	12	0	0	1	2	6.47443
66	45	31	12	0	0	1	2	7.71179
65	44	27	12	0	0	1	2	6.45659
61	41	28	13	0	0	1	2	6.83551
64	39	29	13	0	0	2	2	7.23287
64	45	26	13	0	0	2	2	6.15777
66	45	28	13	0	0	2	2	6.77248
61	38	27	13	0	0	1	2	6.5682
70	34	32	13	0	0	2	2	8.55695
63	43	28	13	0	0	2	1	7.55318
61	42	27	13	0	0	2	2	6.52286
66	44	29	14	0	0	2	2	7.15775
64	42	28	15	0	0	2	2	6.97613
64	9	5	2	1	1	9	0	7.87968
68	27	2	2	1	1	9	0	8.0117
67	15	2	2	1	1	9	0	7.96094

68	15	2	2	1	1	9	0	8.02371
66	14	1	2	1	1	9	0	7.99704
69	9	2	2	1	1	9	0	8.29968
66	18	3	7	0	1	4	0	7.58274
70	9	3	8	1	1	9	2	6.05141
69	25	3	9	0	0	4	0	7.25874
64	10	3	9	0	0	4	0	7.51998
64	37	4	9	0	1	4	0	7.4169
67	45	29	12	0	0	4	2	7.1936
62	44	28	12	0	1	0	0	10.1044
66	44	26	13	0	0	1	2	6.1285
62	38	29	13	0	0	2	2	7.27
64	40	30	13	0	0	5	2	7.81852
61	41	26	13	0	0	1	2	6.17987
74	52	34	13	1	0	1	2	8.84851
76	53	27	13	1	0	1	2	6.80755
68	36	27	14	0	0	2	2	6.79081
66	44	27	14	0	1	1	2	7.4475
68	45	27	15	0	0	0	3	7.41437
60	30	25	12	0	0	4	2	6.31318
56	34	25	13	0	0	0	2	6.05435
51	9	3	2	1	1	9	0	7.27824
64	9	5	1	1	1	9	0	7.95384
43	9	5	1	1	1	9	0	7.12552
59	27	7	1	1	0	9	0	7.56117
68	30	6	1	1	1	9	0	7.91323
71	29	7	1	1	1	9	0	7.97991
66	26	6	1	1	1	9	0	7.75646
60	14	6	1	1	1	9	0	7.5523
60	11	6	1	1	1	9	0	7.64167
69	9	6	1	1	1	9	0	8.29523
60	9	6	1	1	1	9	0	7.71997
59	7	6	1	1	1	9	0	7.75658
54	7	5	1	1	1	9	0	7.49383
54	7	1	2	1	1	9	0	7.53324
47	9	6	2	0	1	9	0	7.27086
47	9	6	2	0	1	9	0	7.27086
47	9	3	3	0	1	9	0	7.23587
37	15	7	3	0	0	9	0	7.04533
37	15	4	4	0	0	9	0	7.03407
66	18	6	4	0	1	9	0	7.57271
70	9	7	4	1	1	9	2	6.15954
66	18	6	6	0	0	3	0	7.45246
51	11	6	6	1	0	3	0	7.07352
38	17	7	6	1	0	9	0	7.08311

66	18	4	6	0	1	9	0	7.55099
70	9	3	6	1	1	9	2	6.0547
66	18	3	7	0	0	3	0	7.48512
51	11	4	7	1	0	3	0	7.04262
51	11	0	7	1	0	3	0	7.26309
67	27	7	9	1	0	9	0	7.35979
42	11	7	9	0	0	9	0	7.35741
59	10	7	9	0	0	9	0	7.61566
48	27	7	9	0	1	9	0	7.4555
50	22	7	9	0	1	9	0	7.39387
58	10	2	9	0	0	4	0	7.25886
45	13	2	9	0	0	4	0	6.85457
69	25	7	10	0	0	9	0	7.38492
59	9	7	10	0	0	9	0	7.8165
45	13	6	11	0	0	9	0	7.56188
64	37	4	8	0	1	9	0	7.60767
42	14	7	6	1	1	9	0	7.27977
58	10	3	8	0	0	9	0	7.43917
62	11	7	10	0	0	9	0	7.7881
59	10	3	8	0	0	9	0	7.47912
37	15	1	6	0	0	9	0	7.18956
66	18	3	8	0	1	9	0	7.63237
45	13	3	10	0	0	9	0	7.34154
42	11	3	10	0	0	9	0	7.41089
64	10	3	10	0	0	9	0	7.88
69	25	3	11	0	0	9	0	7.592
45	13	2	11	0	0	4	0	6.97381
43	9	5	1	1	0	9	0	7.05316
64	23	23	15	0	1	0	1	8.19535
64	37	4	9	0	1	4	0	7.4169
59	40	28	12	0	0	1	2	6.85219
66	44	26	13	0	0	1	2	6.1285
64	40	30	13	0	0	5	2	7.81852
61	41	26	13	0	0	1	2	6.17987
66	44	27	14	0	1	1	2	7.4475
58	3	3	13	0	1	1	3	6.55231

Appendix C: Model Fitting Output (Basic Losses)

Response: Bsc_Loss

Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-5953.441134	Initial	0.10898017	?
2	-2273.809875	Newton	0.07697606	1.61825978
3	-1931.370591	Newton	0.15897145	0.17730285
4	-1888.483874	Newton	0.19300593	0.02270948
5	-1886.524359	Newton	0.04893991	0.00103869
6	-1886.504474	Newton	0.00159458	0.00001054
7	-1886.50446	Newton	0.00000138	7.50003e-9

Converged by Gradient

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	157.0730	18	314.1461	<.0001
Full	1886.5045			
Reduced	2043.5775			

RSquare (U)	0.0769
Observations (or Sum Wgts)	8589

Lack of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack of Fit	7225	1750.7255	3501.451
Pure Error	1345	135.7789	Prob>ChiSq
Total Error	8570	1886.5045	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	3.86763914	0.3517457	120.90	<.0001
PC1	-0.0261601	0.0060243	18.86	<.0001
PC2	-0.0349666	0.0091509	14.60	0.0001
PC3	-0.0800691	0.0179603	19.87	<.0001
PC4	-0.0508744	0.0230839	4.86	0.0275
GENDER[0-1]	-0.0751926	0.0542647	1.92	0.1658
RACE[0-1]	0.18256519	0.0755903	5.83	0.0157
BCF[0-9]	0.02403076	0.1849598	0.02	0.8966
BCF[1-9]	-0.0588789	0.1613642	0.13	0.7152
BCF[2-9]	-0.4080276	0.1543728	6.99	0.0082
BCF[3-9]	-0.1911889	0.1859767	1.06	0.3039
BCF[4-9]	0.21589882	0.2246748	0.92	0.3366
BCF[5-9]	0.29776173	0.458891	0.42	0.5164
BCF[6-9]	-0.0587193	0.2327797	0.06	0.8008
BCF[7-9]	-0.7852953	0.9784122	0.64	0.4222
BCF[8-9]	0.66842924	0.3352942	3.97	0.0462
EDCAT[0-3]	-1.2417504	0.3050132	16.57	<.0001
EDCAT[1-3]	0.62404668	0.2273747	7.53	0.0061
EDCAT[2-3]	0.20081712	0.1594855	1.59	0.2080

Effect Test

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
PC1	1	1	18.856578	0.0000
PC2	1	1	14.600991	0.0001
PC3	1	1	19.874874	0.0000
PC4	1	1	4.857141	0.0275
GENDER	1	1	1.920061	0.1658
RACE	1	1	5.833158	0.0157
BCF	9	9	26.377956	0.0018
EDCAT	3	3	18.877885	0.0003

Appendix D: Model Fitting Output (Policy Losses)

Response: Plcy_Loss
Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-5572.210185	Initial	0.13784189	?
2	-1759.04658	Newton	0.19799154	2.16773221
3	-1216.785829	Newton	2.89629397	0.44564646
4	-1024.612154	Newton	2.92969474	0.18755565
5	-955.4000917	Newton	0.17348585	0.07244226
6	-938.8835216	Newton	0.78898149	0.01759153
7	-937.4324777	Newton	0.16761296	0.00154788
8	-937.3997453	Newton	0.01784804	0.00003492
9	-937.3995557	Newton	0.00015682	0.0000002

Converged by Objective

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	652.7178	11	1305.436	<.0001
Full	937.3996			
Reduced	1590.1174			

RSquare (U)	0.4105
Observations (or Sum Wgts)	8039

Lack of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack of Fit	6369	807.33136	1614.663
Pure Error	1658	130.06820	Prob>ChiSq
Total Error	8027	937.39956	1.0000

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept		14.2287231	0.6186545	528.98	<.0001
PC1		-0.1917909	0.0094847	408.90	<.0001
PC2		-0.0139296	0.0098191	2.01	0.1560
BCF[0-9]		0.33320697	0.2210681	2.27	0.1317
BCF[1-9]		0.28905048	0.226349	1.63	0.2016
BCF[2-9]		-0.3475416	0.2180859	2.54	0.1110
BCF[3-9]		-0.4124425	0.245992	2.81	0.0936
BCF[4-9]		-0.4731062	0.2663767	3.15	0.0757
BCF[5-9]		-0.2431456	0.6015872	0.16	0.6861
BCF[6-9]		0.10247931	0.3796011	0.07	0.7872
BCF[7-9]		0.34812319	1.1429515	0.09	0.7607
BCF[8-9]	Unstable	0.01739157	0.9547189	0.00	0.9855

Effect Test

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
PC1	1	1	408.89548	0.0000
PC2	1	1	2.01249	0.1560
BCF	9	9	31.12071	0.0003

Appendix E: Model Fitting Output (Migrations)

Response: ECF
Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-17589.44753	Initial	5.25307903	?
2	-10363.76468	Newton	22.7559969	0.69720571
3	-445129.1788	Newton	?	?
3	-203682.7946	StepHalve	?	?
3	-84355.62257	StepHalve	?	?
3	-27495.73174	StepHalve	?	?
3	-7626.916955	StepHalve	6.59177607	?
4	-14851.48646	Newton	?	?
4	-6150.010939	StepHalve	0.84927973	?
5	-5623.844245	Newton	0.8602697	0.0935598
6	-5540.76186	Newton	0.3604992	0.01499473
7	-5529.263816	Newton	0.39625045	0.00207949
8	-5527.948909	Newton	0.60012021	0.00023786
9	-5527.764549	Newton	0.60006381	0.00003335
10	-5527.709906	Newton	0.50000851	0.00000989

Converged by Objective

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	9130.950	126	18261.9	0.0000
Full	5527.710			
Reduced	14658.660			

RSquare (U)	0.6229
Observations (or Sum Wgts)	7639

Lack of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack of Fit	57132	5196.4805	10392.96
Pure Error	?	331.2294	Prob>ChiSq
Total Error	7504	5527.7099	1.0000

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	-7.7634803	27.075768	0.08	0.7743
PC1		0.28348471	0.0173209	267.87	<.0001
PC2		-0.0669702	0.0231918	8.34	0.0039
PC3		-1.0439373	0.0635477	269.87	<.0001
PC4		1.00533822	0.065944	232.42	<.0001
GENDER[0-1]		-0.2399141	0.1652359	2.11	0.1465
BCF[0-9]	Unstable	2.40280282	31.301619	0.01	0.9388
BCF[1-9]	Unstable	5.52947784	43.091154	0.02	0.8979
BCF[2-9]	Unstable	-2.3234986	27.071482	0.01	0.9316
BCF[3-9]	Unstable	-2.1636677	27.069495	0.01	0.9363
BCF[4-9]	Unstable	-3.1734805	27.07193	0.01	0.9067
BCF[5-9]	Unstable	4.60362426	84.693631	0.00	0.9567
BCF[6-9]	Unstable	6.41159044	41.226619	0.02	0.8764
BCF[7-9]	Unstable	9.963689	221.95613	0.00	0.9642
BCF[8-9]	Unstable	-3.6346499	33.719852	0.01	0.9142
Intercept	Unstable	0.54479179	39.469869	0.00	0.9890
PC1		0.10776826	0.0158107	46.46	<.0001
PC2		-0.0615533	0.0234734	6.88	0.0087
PC3		-0.4845237	0.0590809	67.26	<.0001
PC4		0.41081229	0.061648	44.41	<.0001

GENDER[0-1]		-0.1413622	0.1660701	0.72	0.3946
BCF[0-9]	Unstable	3.96993947	42.4824	0.01	0.9255
BCF[1-9]	Unstable	9.0106997	51.785995	0.03	0.8619
BCF[2-9]	Unstable	-2.3475884	39.46938	0.00	0.9526
BCF[3-9]	Unstable	-3.3523995	39.468902	0.01	0.9323
BCF[4-9]	Unstable	-3.7413232	39.470749	0.01	0.9245
BCF[5-9]	Unstable	3.46021801	89.435258	0.00	0.9691
BCF[6-9]	Unstable	7.22306466	50.24137	0.02	0.8857
BCF[7-9]	Unstable	-3.0019124	339.97883	0.00	0.9930
BCF[8-9]	Unstable	6.37942297	42.106414	0.02	0.8796
Intercept	Unstable	4.4010041	38.217553	0.01	0.9083
PC1		0.06319679	0.0144266	19.19	<.0001
PC2		-0.0067879	0.0220267	0.09	0.7580
PC3		-0.4120522	0.055532	55.06	<.0001
PC4		0.35532912	0.0559703	40.30	<.0001
GENDER[0-1]		0.24073907	0.1599325	2.27	0.1323
BCF[0-9]	Unstable	2.15294376	41.32256	0.00	0.9584
BCF[1-9]	Unstable	3.54584336	50.839128	0.00	0.9444
BCF[2-9]	Unstable	-0.210329	38.217908	0.00	0.9956
BCF[3-9]	Unstable	-4.2923647	38.216837	0.01	0.9106
BCF[4-9]	Unstable	-3.9150635	38.218275	0.01	0.9184
BCF[5-9]	Unstable	3.73747473	88.886351	0.00	0.9665
BCF[6-9]	Unstable	6.21857871	49.264245	0.02	0.8996
BCF[7-9]	Unstable	-3.8992633	330.57625	0.00	0.9906
BCF[8-9]	Unstable	4.98359396	40.935829	0.01	0.9031
Intercept	Unstable	2.19119335	40.317579	0.00	0.9567
PC1		0.07829869	0.0153472	26.03	<.0001
PC2		-0.0427199	0.0233579	3.34	0.0674
PC3		-0.4187843	0.0584639	51.31	<.0001
PC4		0.22230613	0.0606906	13.42	0.0002
GENDER[0-1]		-0.1879751	0.1653033	1.29	0.2555
BCF[0-9]	Unstable	2.59711228	43.271758	0.00	0.9521
BCF[1-9]	Unstable	4.38193062	52.435873	0.01	0.9334
BCF[2-9]	Unstable	-3.1315854	40.317814	0.01	0.9381
BCF[3-9]	Unstable	0.61438062	40.315678	0.00	0.9878
BCF[4-9]	Unstable	-2.2102462	40.317467	0.00	0.9563
BCF[5-9]	Unstable	3.32440781	89.812721	0.00	0.9705
BCF[6-9]	Unstable	7.86187687	50.910081	0.02	0.8773
BCF[7-9]	Unstable	-4.6308116	350.18967	0.00	0.9894
BCF[8-9]	Unstable	1.71601903	42.911618	0.00	0.9681
Intercept	Unstable	2.79097684	39.227725	0.01	0.9433
PC1		0.07093739	0.0130442	29.57	<.0001
PC2		-0.0440028	0.0193085	5.19	0.0227
PC3		-0.3916261	0.0517257	57.32	<.0001
PC4		0.18361994	0.0496958	13.65	0.0002
GENDER[0-1]		0.18228687	0.1510289	1.46	0.2274
BCF[0-9]	Unstable	1.80307582	42.259577	0.00	0.9660
BCF[1-9]	Unstable	2.47738708	51.605805	0.00	0.9617
BCF[2-9]	Unstable	-3.2390536	39.229389	0.01	0.9342
BCF[3-9]	Unstable	-2.8571281	39.227633	0.01	0.9419
BCF[4-9]	Unstable	-0.2566833	39.228909	0.00	0.9948
BCF[5-9]	Unstable	3.08957027	89.329461	0.00	0.9724
BCF[6-9]	Unstable	6.8676714	50.052628	0.02	0.8909
BCF[7-9]	Unstable	-4.6858496	340.03194	0.00	0.9890
BCF[8-9]	Unstable	3.25046415	41.882688	0.01	0.9381
Intercept	Unstable	0.23941071	38.858812	0.00	0.9951
PC1		0.10388396	0.0227193	20.91	<.0001
PC2		-0.0697711	0.0325602	4.59	0.0321
PC3		-0.5087317	0.0787117	41.77	<.0001

PC4		0.44290725	0.0863892	26.28	<.0001
GENDER[0-1]		-0.1577218	0.209157	0.57	0.4508
BCF[0-9]	Unstable	1.77521564	41.909664	0.00	0.9662
BCF[1-9]	Unstable	2.20602027	51.324051	0.00	0.9657
BCF[2-9]	Unstable	-3.4898843	38.852146	0.01	0.9284
BCF[3-9]	Unstable	-4.1204748	38.853979	0.01	0.9155
BCF[4-9]	Unstable	-2.6714302	38.851428	0.00	0.9452
BCF[5-9]	Unstable	8.01621692	89.159503	0.01	0.9284
BCF[6-9]	Unstable	5.48578606	49.760845	0.01	0.9122
BCF[7-9]	Unstable	-1.9628989	336.49176	0.00	0.9953
BCF[8-9]	Unstable	4.26649997	41.528872	0.01	0.9182
Intercept	Unstable	2.01019754	34.483887	0.00	0.9535
PC1		0.00005185	0.0172785	0.00	0.9976
PC2		0.01963546	0.0298648	0.43	0.5109
PC3		-0.2456325	0.0635869	14.92	0.0001
PC4		-0.2225706	0.0778105	8.18	0.0042
GENDER[0-1]		-0.910182	0.2068778	19.36	<.0001
BCF[0-9]	Unstable	-2.5051309	40.132571	0.00	0.9502
BCF[1-9]	Unstable	5.71917173	48.098551	0.01	0.9054
BCF[2-9]	Unstable	-1.267131	34.484906	0.00	0.9707
BCF[3-9]	Unstable	-2.2668141	34.484943	0.00	0.9476
BCF[4-9]	Unstable	-2.6514205	34.488673	0.01	0.9387
BCF[5-9]	Unstable	-2.2805523	107.08522	0.00	0.9830
BCF[6-9]	Unstable	11.1151348	46.426481	0.06	0.8108
BCF[7-9]	Unstable	-3.5005266	288.56867	0.00	0.9903
BCF[8-9]	Unstable	4.90739965	37.476293	0.02	0.8958
Intercept		-9.8161605	29.243737	0.11	0.7371
PC1		0.17407585	0.0944884	3.39	0.0654
PC2		-0.4459437	0.1275202	12.23	0.0005
PC3		-0.8186496	0.2182547	14.07	0.0002
PC4		0.86658661	0.2980992	8.45	0.0036
GENDER[0-1]		-2.1432972	1.4092494	2.31	0.1283
BCF[0-9]		-1.9030057	35.082555	0.00	0.9567
BCF[1-9]		-1.2097039	53.88266	0.00	0.9821
BCF[2-9]		-1.7960089	28.87632	0.00	0.9504
BCF[3-9]		-0.9149753	28.872719	0.00	0.9747
BCF[4-9]	Unstable	-8.7852323	45.491182	0.04	0.8469
BCF[5-9]		-1.5451075	113.83574	0.00	0.9892
BCF[6-9]	Unstable	9.93083548	42.429614	0.05	0.8149
BCF[7-9]	Unstable	17.8037689	222.18792	0.01	0.9361
BCF[8-9]	Unstable	0.07514688	35.240255	0.00	0.9983
Intercept	Unstable	2.97415936	39.09437	0.01	0.9394
PC1		0.03943366	0.0169501	5.41	0.0200
PC2		-0.0837163	0.0292441	8.19	0.0042
PC3		-0.330621	0.0613927	29.00	<.0001
PC4		0.2829612	0.0676572	17.49	<.0001
GENDER[0-1]		0.05765936	0.1724015	0.11	0.7380
BCF[0-9]	Unstable	1.57101409	42.136311	0.00	0.9703
BCF[1-9]	Unstable	5.97326457	51.500097	0.01	0.9077
BCF[2-9]	Unstable	-1.5555554	39.093097	0.00	0.9683
BCF[3-9]	Unstable	-3.4058328	39.093196	0.01	0.9306
BCF[4-9]	Unstable	-3.0109195	39.094244	0.01	0.9386
BCF[5-9]	Unstable	5.61802214	89.266416	0.00	0.9498
BCF[6-9]	Unstable	6.4725715	49.946583	0.02	0.8969
BCF[7-9]	Unstable	-1.8352884	336.27577	0.00	0.9956
BCF[8-9]	Unstable	7.54104622	41.753882	0.03	0.8567

Effect Test

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq

PC1	9	9	512.5851	0.0000
PC2	9	9	56.5793	0.0000
PC3	9	9	445.4168	0.0000
PC4	9	9	432.5197	0.0000
GENDER	9	9	127.3318	0.0000
BCF	81	81	5166.2422	0.0000

Appendix F: Model Fitting Output (Promotions)

Response: EGRADE

Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-21179.80525	Initial	0.11696276	?
2	-13874.82838	Newton	0.28240041	0.52649095
3	-12534.84159	Newton	2.70201914	0.10690089
4	-8889.897517	Newton	8.46109098	0.41000922
5	-6883.662008	Newton	60.9869467	0.29144845
6	-5660.677684	Newton	86.4959241	0.21604872
7	-4951.685799	Newton	103.673582	0.14318163
8	-4592.393135	Newton	103.681409	0.07823631
9	-4449.866492	Newton	88.9090165	0.03202935
10	-4403.909568	Newton	66.6723805	0.01043546
11	-4391.71218	Newton	40.887586	0.00277736
12	-4389.244608	Newton	22.9178859	0.00056218
13	-4388.803229	Newton	16.1888128	0.00010057
14	-4388.686843	Newton	15.1023992	0.00002652
15	-4388.645265	Newton	15.0284976	0.00000947

Converged by Objective

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	13030.906	90	26061.81	0.0000
Full	4388.645			
Reduced	17419.552			

RSquare (U)	0.7481
Observations (or Sum Wgts)	7639

Lack of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack of Fit	83445	4046.3211	8092.642
Pure Error	?	342.3242	Prob>ChiSq
Total Error	7534	4388.6453	1.0000

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	676.522449	99.770496	45.98	<.0001
PC1	Unstable	-13.970984	1.50788	85.85	<.0001
PC2	Unstable	0.96447898	0.3906626	6.10	0.0136
PC3	Unstable	44.1492679	4.5325053	94.88	<.0001
PC4	Unstable	-41.034798	4.8156301	72.61	<.0001
GENDER[0-1]		-0.8510603	0.8188937	1.08	0.2987
RACE[0-1]		0.9930271	1.0364288	0.92	0.3380
Intercept	Unstable	672.879197	99.676597	45.57	<.0001
PC1	Unstable	-13.671197	1.4828449	85.00	<.0001
PC2	Unstable	1.20020843	0.2162626	30.80	<.0001
PC3	Unstable	43.8292531	4.495531	95.05	<.0001
PC4	Unstable	-40.462588	4.7517065	72.51	<.0001
GENDER[0-1]		-0.9127247	0.6553992	1.94	0.1637
RACE[0-1]		0.80515398	0.8507255	0.90	0.3439
Intercept	Unstable	675.195666	99.668673	45.89	<.0001
PC1	Unstable	-13.435138	1.4793724	82.48	<.0001
PC2	Unstable	0.91130017	0.2328542	15.32	<.0001
PC3	Unstable	42.8301461	4.4855392	91.17	<.0001
PC4	Unstable	-38.855632	4.738471	67.24	<.0001
GENDER[0-1]		-1.2192532	0.6065891	4.04	0.0444

RACE[0-1]		0.31037332	0.7946855	0.15	0.6961
Intercept	Unstable	667.313111	99.657676	44.84	<.0001
PC1	Unstable	-12.781526	1.4770748	74.88	<.0001
PC2	Unstable	1.17037307	0.2019528	33.59	<.0001
PC3	Unstable	40.979123	4.4790707	83.70	<.0001
PC4	Unstable	-37.231744	4.7326461	61.89	<.0001
GENDER[0-1]		-1.3109525	0.5813392	5.09	0.0241
RACE[0-1]		0.30923923	0.7703843	0.16	0.6881
Intercept	Unstable	658.839181	99.650407	43.71	<.0001
PC1	Unstable	-12.143587	1.4756949	67.72	<.0001
PC2	Unstable	1.18941486	0.194882	37.25	<.0001
PC3	Unstable	38.6338162	4.4740739	74.56	<.0001
PC4	Unstable	-35.228993	4.7294648	55.49	<.0001
GENDER[0-1]		-1.3910589	0.567336	6.01	0.0142
RACE[0-1]		0.42132294	0.7566046	0.31	0.5776
Intercept	Unstable	646.008152	99.643865	42.03	<.0001
PC1	Unstable	-11.553048	1.4747566	61.37	<.0001
PC2	Unstable	1.31448251	0.1884	48.68	<.0001
PC3	Unstable	36.8812612	4.4711266	68.04	<.0001
PC4	Unstable	-34.214742	4.7275684	52.38	<.0001
GENDER[0-1]		-0.7276689	0.5599529	1.69	0.1938
RACE[0-1]		0.16777188	0.7492339	0.05	0.8228
Intercept	Unstable	634.385714	99.636668	40.54	<.0001
PC1	Unstable	-11.015789	1.4738993	55.86	<.0001
PC2	Unstable	1.29558861	0.1857029	48.67	<.0001
PC3	Unstable	35.0118796	4.4680297	61.40	<.0001
PC4	Unstable	-32.42416	4.7244931	47.10	<.0001
GENDER[0-1]		-0.969244	0.5495034	3.11	0.0778
RACE[0-1]		0.59656537	0.7383815	0.65	0.4191
Intercept	Unstable	620.718201	99.63458	38.81	<.0001
PC1	Unstable	-10.60855	1.4737295	51.82	<.0001
PC2	Unstable	1.26407249	0.185552	46.41	<.0001
PC3	Unstable	33.6833537	4.4674959	56.85	<.0001
PC4	Unstable	-31.441536	4.7243309	44.29	<.0001
GENDER[0-1]		-0.6701116	0.5521817	1.47	0.2249
RACE[0-1]		0.49950695	0.74046	0.46	0.4999
Intercept	Unstable	609.408717	99.62141	37.42	<.0001
PC1	Unstable	-10.196132	1.4728631	47.92	<.0001
PC2	Unstable	1.19510817	0.1838965	42.23	<.0001
PC3	Unstable	32.4451454	4.4646713	52.81	<.0001
PC4	Unstable	-29.765764	4.7210973	39.75	<.0001
GENDER[0-1]		-0.6636778	0.5424615	1.50	0.2212
RACE[0-1]		0.64287406	0.7285671	0.78	0.3776
Intercept	Unstable	591.294863	99.630159	35.22	<.0001
PC1	Unstable	-9.782685	1.4731916	44.10	<.0001
PC2	Unstable	1.19201321	0.1846417	41.68	<.0001
PC3	Unstable	31.0232869	4.4658017	48.26	<.0001
PC4	Unstable	-28.661473	4.7222925	36.84	<.0001
GENDER[0-1]		-0.301459	0.5525976	0.30	0.5854
RACE[0-1]		0.62837881	0.7413493	0.72	0.3967
Intercept	Unstable	585.790591	99.611355	34.58	<.0001
PC1	Unstable	-9.5842897	1.4724453	42.37	<.0001
PC2	Unstable	1.10783039	0.182878	36.70	<.0001
PC3	Unstable	30.3936386	4.4632158	46.37	<.0001
PC4	Unstable	-27.691363	4.7197748	34.42	<.0001
GENDER[0-1]		-0.4352369	0.5365453	0.66	0.4173
RACE[0-1]		0.75811113	0.7190128	1.11	0.2917
Intercept	Unstable	540.216085	99.557947	29.44	<.0001
PC1	Unstable	-8.6434445	1.4709754	34.53	<.0001

PC2	Unstable	1.04361589	0.1819174	32.91	<.0001
PC3	Unstable	27.2086815	4.4582558	37.25	<.0001
PC4	Unstable	-24.870447	4.7154475	27.82	<.0001
GENDER[0-1]		-0.0290267	0.5316484	0.00	0.9565
RACE[0-1]		0.70255024	0.7039719	1.00	0.3183
Intercept	Unstable	437.472977	99.109047	19.48	<.0001
PC1	Unstable	-6.7600445	1.4607528	21.42	<.0001
PC2	Unstable	0.8337319	0.1804241	21.35	<.0001
PC3	Unstable	21.0931596	4.4257388	22.71	<.0001
PC4	Unstable	-19.104555	4.6830877	16.64	<.0001
GENDER[0-1]		0.27148665	0.5269446	0.27	0.6064
RACE[0-1]		1.09118284	0.6894621	2.50	0.1135
Intercept	Unstable	330.858489	98.806367	11.21	0.0008
PC1	Unstable	-4.9671808	1.4548915	11.66	0.0006
PC2	Unstable	0.66339663	0.1790251	13.73	0.0002
PC3	Unstable	15.109373	4.4061737	11.76	0.0006
PC4	Unstable	-13.931697	4.6652738	8.92	0.0028
GENDER[0-1]		0.478587	0.5171772	0.86	0.3548
RACE[0-1]		0.51270559	0.6565841	0.61	0.4349
Intercept	Unstable	174.882926	97.410716	3.22	0.0726
PC1	Unstable	-2.5436063	1.4317081	3.16	0.0756
PC2	Unstable	0.34506713	0.173966	3.93	0.0473
PC3	Unstable	7.28946313	4.3300131	2.83	0.0923
PC4	Unstable	-7.0694449	4.5945789	2.37	0.1239
GENDER[0-1]		0.94027732	0.4812227	3.82	0.0507
RACE[0-1]		-0.0266745	0.5835874	0.00	0.9635

Effect Test

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
PC1	15	15	1559.0124	0.0000
PC2	15	15	264.3712	0.0000
PC3	15	15	1657.0780	0.0000
PC4	15	15	1482.1123	0.0000
GENDER	15	15	117.9691	0.0000
RACE	15	15	34.9552	0.0025

Appendix G: Model Fitting Output (New Hires)

Response: New_Hires

Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-5622.116782	Initial	0.11358345	?
11	-1312.861339	Newton	0.87484645	0.00000462

Converged by Objective

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	487.4976	29	974.9952	<.0001
Full	1312.8613			
Reduced	1800.3589			

RSquare (U)	0.2708
Observations (or Sum Wgts)	8111

Lack of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack of Fit	431	237.4106	474.8211
Pure Error	7650	1075.4508	Prob>ChiSq
Total Error	8081	1312.8613	0.0711

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	5.30800205	10.399575	0.26	0.6098
EGRADE[1-16]	Unstable	-0.4071445	104.80681	0.00	0.9969
EGRADE[2-16]	Unstable	-7.7727628	7.1615447	1.18	0.2778
EGRADE[3-16]	Unstable	-0.5596703	7.1388915	0.01	0.9375
EGRADE[4-16]	Unstable	-2.1789283	7.1101729	0.09	0.7593
EGRADE[5-16]	Unstable	-1.5248986	7.1077123	0.05	0.8301
EGRADE[6-16]	Unstable	0.36163418	7.1276348	0.00	0.9595
EGRADE[7-16]	Unstable	-1.8271008	7.10699	0.07	0.7971
EGRADE[8-16]	Unstable	7.22550786	20.513454	0.12	0.7247
EGRADE[9-16]	Unstable	-1.3164623	7.1069673	0.03	0.8530
EGRADE[10-16]	Unstable	-1.344349	7.1721176	0.04	0.8513
EGRADE[11-16]	Unstable	0.65605727	7.1085672	0.01	0.9265
EGRADE[12-16]	Unstable	0.32137698	7.1073495	0.00	0.9639
EGRADE[13-16]	Unstable	1.31274891	7.108882	0.03	0.8535
EGRADE[14-16]	Unstable	2.02581357	7.1132391	0.08	0.7758
EGRADE[15-16]	Unstable	2.48396953	7.1232463	0.12	0.7273
GENDER[0-1]		-0.5093977	0.0603864	71.16	<.0001
RACE[0-1]		0.18152776	0.0792194	5.25	0.0219
ECF[0-9]		0.23344282	0.5122632	0.21	0.6486
ECF[1-9]		-1.3112864	0.2267569	33.44	<.0001
ECF[2-9]		-1.4563797	0.2102151	48.00	<.0001
ECF[3-9]		-1.0724661	0.2326672	21.25	<.0001
ECF[4-9]		0.39544902	0.3326242	1.41	0.2345
ECF[5-9]		-0.2973376	0.6836192	0.19	0.6636
ECF[6-9]		-0.4134469	0.2821575	2.15	0.1428
ECF[7-9]		-2.3946162	0.703384	11.59	0.0007
ECF[8-9]		-0.6960104	0.2688067	6.70	0.0096
EDCAT[0-3]	Unstable	6.26235034	22.782047	0.08	0.7834
EDCAT[1-3]	Unstable	-0.5721114	7.5977081	0.01	0.9400
EDCAT[2-3]	Unstable	-1.12333	7.5947092	0.02	0.8824

Effect Test

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
EGRADE	15	15	386.75057	0.0000
GENDER	1	1	71.15991	0.0000
RACE	1	1	5.25077	0.0219
ECF	9	9	91.25817	0.0000
EDCAT	3	3	160.49378	0.0000

Appendix H: Visual Basic Code

```
' This model was developed by Stephen G Hoffman.
'
' The customer's first step in running this model is to
' collect the most recent fiscal year inventory snapshot
' from the EDPSMSTR file.  A SAS program (PREP.SAS -
' provided with the model) will accept the EDPSMSTR fields
' and output usable variables.  Once the records have been
' verified they may be copied into the appropriate cells on
' the BI worksheet.  The following subroutines may then be
' run to produce and manipulate single or multiple-year
' forecasts of the NSA's civilian personnel inventory.
' After ten forecast years the model must be modified to
' pull-down farther.

Sub PullDown()
Dim Flag As Integer, LastRow As Integer, RelativeLastRow As
Integer
Dim CopyBlock As String, PivotBlock As String, SumBlock As
String

    Application.DisplayStatusBar = True
    Application.ScreenUpdating = False

'   Display percent completed in status bar
    Application.StatusBar = "Working ... "

'   Go to Beginning Inventory sheet and find bottom of data
    Sheets("BI").Select
    Flag = 1
    LastRow = 5
    Do While Flag = 1
     If Cells(LastRow, 2) <> "" Then
       LastRow = LastRow + 1
     Else
       Flag = 0
     End If
    Loop
    LastRow = LastRow + 17

    CopyBlock = "BI_PC1:N" & LTrim(Str(LastRow))
    Worksheets("BI").Range("BI_PC1:BI_PC4"). _
    Copy (Worksheets("BI").Range(CopyBlock))

'   Pull down Basic Loss formulas
    CopyBlock = "BLSI_LOGIT:T" & LTrim(Str(LastRow))
    Worksheets("BLSI").Range("BLSI_LOGIT:BLSI_PC4"). _
    Copy (Worksheets("BLSI").Range(CopyBlock))
```

```
    Application.StatusBar = "Working ...    Basic Losses
Completed"

'   Pull down Policy Loss formulas
    CopyBlock = "PLSI_LOGIT:T" & LTrim(Str(LastRow))
    Worksheets("PLSI").Range("PLSI_LOGIT:PLSI_PC4"). _
    Copy (Worksheets("PLSI").Range(CopyBlock))
    Application.StatusBar = "Working ...    Policy Losses
Completed"

'   Pull down Migrate formulas
    CopyBlock = "MI_LOGIT:AB" & LTrim(Str(LastRow))
    Worksheets("MI").Range("MI_LOGIT:MI_PC4"). _
    Copy (Worksheets("MI").Range(CopyBlock))
    Application.StatusBar = "Working ...    Migrations
Completed"

'   Pull down Promote formulas
    CopyBlock = "PI_LOGIT:AH" & LTrim(Str(LastRow))
    Worksheets("PI").Range("PI_LOGIT:PI_PC4"). _
    Copy (Worksheets("PI").Range(CopyBlock))
    Application.StatusBar = "Working ...    Promotions
Completed"

'   Pull down New Hires formulas
    CopyBlock = "NH_LOGIT:Q" & LTrim(Str(LastRow))
    Worksheets("NH").Range("NH_LOGIT:NH_EDCAT"). _
    Copy (Worksheets("NH").Range(CopyBlock))
    Application.StatusBar = "Working ...    New Hires
Completed"

'   Pull down Ending Inventory formulas
    LastRow = LastRow + 2
    CopyBlock = "EI_INV:J" & LTrim(Str(LastRow))
    Worksheets("EI").Range("EI_INV:EI_EDCAT"). _
    Copy (Worksheets("EI").Range(CopyBlock))

'   Add New Hires to EI
    RelativeLastRow = LastRow - 6
    SumBlock = "=SUM(NH!RC[7]:R[" & RelativeLastRow &
"]C[7])*.6"
    Range("EI_NH0").Value = SumBlock
    RelativeLastRow = LastRow - 7
    SumBlock = "=SUM(NH!R[-1]C[7]:R[" & RelativeLastRow &
"]C[7])*.4"
    Range("EI_NH1").Value = SumBlock
    Application.StatusBar = "Working ...    Ending Inventory
Completed"
```

```
'    Set forecast year counter to one
     Sheets("SP").Select
     Range("G1").Value = "=1"

'    Insert shaded box for totals
     Range("A3:BP37").Delete Shift:=xlUp

Range("A3,B3,C3,D3,E3,F3,A4,B4,C4,D4,E4,F4").BorderAround _
     Weight:=xlThin, ColorIndex:=xlAutomatic
     Range("A3:F4").BorderAround Weight:=xlMedium,
ColorIndex:=xlAutomatic
     Range("A4,A3:F3").Select
     With Selection.Interior
          .ColorIndex = 40
          .Pattern = xlSolid
     End With

'    Calculate Grand Totals
     Range("A4").Value = "Grand Total"
     Range("B3").Value = "Beg Inv"
     Range("C3").Value = "Basic Losses"
     Range("D3").Value = "Policy Losses"
     Range("E3").Value = "New Hires"
     Range("F3").Value = "Ending Inv"

     RelativeLastRow = LastRow - 6
     SumBlock = "=SUM(BI!RC:R[" & RelativeLastRow & "]C)"
     Range("B4").Value = SumBlock
     SumBlock = "=SUM(BLSI!RC[5]:R[" & RelativeLastRow &
"]C[5])-SP!RC[-1]"
     Range("C4").Value = SumBlock
     SumBlock = "=SUM(PLSI!RC[4]:R[" & RelativeLastRow _
     & "]C[4])-SUM(BLSI!RC[4]:R[" & RelativeLastRow &
"]C[4])"
     Range("D4").Value = SumBlock
     SumBlock = "=SUM(NH!RC[4]:R[" & RelativeLastRow &
"]C[4])"
     Range("E4").Value = SumBlock
     RelativeLastRow = LastRow - 4
     SumBlock = "=SUM(EI!RC[-4]:R[" & RelativeLastRow & "]C[-
4])"
     Range("F4").Value = SumBlock

'    Create Pivot Table to Aggregate GENDER in Scratch Pad
     Range("A6").Select
     PivotBlock = "EI!R3C2:R" & LTrim(Str(LastRow)) & "C10"
     ActiveSheet.PivotTableWizard SourceType:=xlDatabase,
SourceData:=
          PivotBlock, TableDestination:="R6C1", TableName:= _
```

```
        "GndrTable", RowGrand:=False, ColumnGrand:=False
    ActiveSheet.PivotTables("GndrTable").AddFields
ColumnFields:="GENDER"

ActiveSheet.PivotTables("GndrTable").PivotFields("EI").Orien
tation = xlDataField

'    Create Pivot Table to Aggregate RACE in Scratch Pad
    Range("A10").Select
    ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R10C1",
TableName:= "RaceTable", RowGrand:=False, ColumnGrand:=False
    ActiveSheet.PivotTables("RaceTable").AddFields
ColumnFields:="RACE"

ActiveSheet.PivotTables("RaceTable").PivotFields("EI").Orien
tation = xlDataField

'    Create Pivot Table to Aggregate EDCAT in Scratch Pad
    Range("A14").Select
    ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R14C1",
TableName:= "EdcatTable", RowGrand:=False,
ColumnGrand:=False
    ActiveSheet.PivotTables("EdcatTable").AddFields
ColumnFields:="EDCAT"

ActiveSheet.PivotTables("EdcatTable").PivotFields("EI").Orie
ntation = xlDataField

'    Create Pivot Table to Aggregate CF in Scratch Pad
    Range("A18").Select
    ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R18C1",
TableName:= "CFTable", RowGrand:=False, ColumnGrand:=False
    ActiveSheet.PivotTables("CFTable").AddFields
ColumnFields:= "CF"

ActiveSheet.PivotTables("CFTable").PivotFields("EI").Orienta
tion = xlDataField

'    Create Pivot Table to Aggregate GRADE in Scratch Pad
    Range("A22").Select
    ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R22C1",
TableName:= "GrdTable", RowGrand:=False, ColumnGrand:=False
    ActiveSheet.PivotTables("GrdTable").AddFields
ColumnFields:= "GRADE"
```

```
ActiveSheet.PivotTables("GrdTable").PivotFields("EI").Orient
ation = xlDataField

'    Create Pivot Table to Aggregate TIG in Scratch Pad
     Range("A26").Select
     ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R26C1",
TableName:= "TigTable", RowGrand:=False, ColumnGrand:=False
     ActiveSheet.PivotTables("TigTable").AddFields
ColumnFields:="TIG"

ActiveSheet.PivotTables("TigTable").PivotFields("EI").Orient
ation = xlDataField

'    Create Pivot Table to Aggregate AGE in Scratch Pad
     Range("A30").Select
     ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R30C1",
TableName:= "AgeTable", RowGrand:=False, ColumnGrand:=False
     ActiveSheet.PivotTables("AgeTable").AddFields
ColumnFields:= "AGE"

ActiveSheet.PivotTables("AgeTable").PivotFields("EI").Orient
ation = xlDataField

'    Create Pivot Table to Aggregate YOS in Scratch Pad
     Range("A34").Select
     ActiveSheet.PivotTableWizard SourceType:=xlPivotTable,
SourceData:= "GndrTable", TableDestination:="R34C1",
TableName:= "YosTable", RowGrand:=False, ColumnGrand:=False
     ActiveSheet.PivotTables("YosTable").AddFields
ColumnFields:= "YOS"

ActiveSheet.PivotTables("YosTable").PivotFields("EI").Orient
ation = xlDataField

'    Modify SP cell format
     Columns("A:BP").NumberFormat = "0"
     Columns("A:BP").EntireColumn.AutoFit
     Range("A1").Select
     Application.StatusBar = "Working ...    Scratch Pad
Completed"

'    Return prompt to Beginning Inventory sheet
     Sheets("BI").Select
     Range("A1").Select
     Beep
```

```
'    Reset status bar
     Application.StatusBar = False

End Sub


Sub Refresh()

     Application.ScreenUpdating = False
     Sheets("SP").Select
     Range("A6").Select
     ActiveSheet.PivotTables("GndrTable").RefreshTable
     Range("A10").Select
     ActiveSheet.PivotTables("RaceTable").RefreshTable
     Range("A14").Select
     ActiveSheet.PivotTables("EdcatTable").RefreshTable
     Range("A18").Select
     ActiveSheet.PivotTables("CFTable").RefreshTable
     Range("A22").Select
     ActiveSheet.PivotTables("GrdTable").RefreshTable
     Range("A26").Select
     ActiveSheet.PivotTables("TigTable").RefreshTable
     Range("A30").Select
     ActiveSheet.PivotTables("AgeTable").RefreshTable
     Range("A34").Select
     ActiveSheet.PivotTables("YosTable").RefreshTable

     Columns("A:BP").NumberFormat = "0"
     Columns("A:BP").EntireColumn.AutoFit
     Range("A1").Select

End Sub


Sub ForecastAnother()
Dim Flag As Integer, LastRow As Integer, AddOne As Integer
Dim CopyBlock As String, NextFcst As String

     Application.DisplayStatusBar = True
     Application.ScreenUpdating = False

'    Display percent completed in status bar
     Application.StatusBar = "Working ... "

'    Go to Ending Inventory sheet and find bottom of data
     Sheets("EI").Select
     Flag = 1
     LastRow = 5
     Do While Flag = 1
```

```
      If Cells(LastRow, 9) <> "" Then
        LastRow = LastRow + 1
      Else
        Flag = 0
      End If
    Loop
    LastRow = LastRow - 1

'   Copy EI to BI for future forecast
    CopyBlock = "EI_NH0:J" & LTrim(Str(LastRow))
    Range(CopyBlock).Select
    Selection.Copy
    Sheets("BI").Select
    Range("BI_INV").Select
    Selection.PasteSpecial Paste:=xlValues,
Operation:=xlNone, _
        SkipBlanks:=False, Transpose:=False
    Application.CutCopyMode = False

    Selection.NumberFormat = "0"
    Range("A1").Select
    Sheets("EI").Select
    Range("A1").Select
    Application.StatusBar = "Working ...    Ending Inventory
Completed"

'   Re-set Basic Loss control lever to zero
    ActiveWorkbook.Names.Add Name:="BLCtrLvr", _
    RefersToR1C1:="=0"

'   Re-set Policy Loss control lever to zero
    ActiveWorkbook.Names.Add Name:="PLCtrLvr", _
    RefersToR1C1:="=0"

'   Re-set New Hires control lever to zero
    ActiveWorkbook.Names.Add Name:="NHCtrLvr", _
    RefersToR1C1:="=0"

    Call Refresh

'   Set forecast year counter to next
    Worksheets("SP").Range("G1").Value =
    Worksheets("SP").Range("G1").Value + 1
    Beep

'   Reset status bar
    Application.StatusBar = False

End Sub
```

Bibliography

Agresti, A. *Categorical Data Analysis*. New York: John Wiley & Sons, 1990.

Christensen, R. *Log-Linear Models*. New York: Springer-Verlag, 1990.

Clinton, W. J., President of the United States. The President's Statement on the Buyout Program for Federal Employees. April 4, 1995.

Department of Defense. *Quadrennial Defense Review*. Washington: GPO, 1996.

Dillon, W. R. and M Goldstein. *Multivariate Analysis*. New York: John Wiley & Sons, 1984.

JMP® Statistics and Graphics Guide, Version 3.1, (JMP 3.2 Professional Edition - What's New!), Cary: SAS Institute, 1995.

Law, A. M. and W. D. Kelton. *Simulation Modeling and Analysis*. New York: McGraw-Hill, Inc., 1991.

Makridakis, S. and S. C. Wheelwright and V. E. McGee. *Forecasting: Methods and Applications* (Second Edition). New York: John Wiley & Sons, 1983.

McCullagh, P. and J. A. Nelder. *Generalized Linear Models* (Second Edition). New York: Chapman and Hall, 1989.

RAND Corporation, A Project Air Force Report by Warren E. Walker and The Enlisted Force Management Project Team, *Design and Development of an Enlisted Force Management System for the Air Force*. Santa Monica, CA, 1991.

U. S. Bureau of the Census, Table No. 527. Federal Civilian Employment, by Branch and Agency: 1980 to 1992. *Statistical Abstract of the United States*: 1994 (114th Edition). Washington, DC, 1994.

U. S. Bureau of the Census, Table No. 527. Federal Civilian Employment, by Branch and Agency: 1980 to 1996. *Statistical Abstract of the United States*: 1996 (116th Edition). Washington, DC, 1996.

Woolsey, R. E. D. and H. F. Swanson. *Operations Research for Intermediate Application: A Quick and Dirty Manual*. New York: Harper & Row, 1975.

Yan, T. W. *Binary Logistic Regression With Data That Have No Variance On The Dependent Variable: An Application To College Dropout Analysis*. Ph.D. dissertation. University of California, Los Angeles CA, 1992.

Vita

Captain Stephen G. Hoffman was born [redacted] in [redacted]. He graduated from Ottawa-Glandorf High School in 1984 and enlisted in the Air Force on October 31st of that same year. His first assignment was at Warner Robins AFB as a KC/EC-135 aircraft mechanic. While at Warner Robins AFB he earned an Associate's Degree from Macon College. On 7 December 1991 he graduated from Valdosta State University with a Bachelor's Degree in Applied Mathematics and was commissioned, but wasn't called to Active Duty until the following October. He continued at VSU in the interim and earned a Master's Degree in Public Administration.

His second assignment was at Bolling AFB, DC, as a Personnel Analyst for the Air Force Personnel Operations Agency. Captain Hoffman remained at Bolling until his selection to the Air Force Institute of Technology, School of Engineering, Department of Operational Sciences in 1996. He has been assigned to the Air Force Operational Test and Evaluation Center at Kirtland AFB, NM, where he will work in the Logistics and Support Systems Division.

Captain Hoffman and his wife, the former Karen Michelle Mizell of Warner Robins, Georgia, have two children: Jessica and Joshua.

Permanent address: [redacted]

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1998	Master's Thesis

4. TITLE AND SUBTITLE
A Model to Forecast Civilian Personnel Inventory for the National Security Agency

5. FUNDING NUMBERS

6. AUTHOR(S)
Stephen G. Hoffman, Captain, USAF

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

Air Force Institute of Technology
Wright-Patterson AFB, OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/98M-13

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Security Agency Workforce Assessment Team
9800 Savage Road
Fort George Meade, MD 20755-6000
Attn: Sue Rose or Mary Filippell (S223)
(410) 859-6426

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Dr. Kenneth W. Bauer Jr.
255-6565 x4326 (work) 656-4943 (fax)
kbauer@afit.af.mil

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this thesis is to develop a model to forecast civilian personnel inventory for the National Security Agency (NSA). Accurate prediction of personnel inventories will help the NSA with issues ranging from hiring, promotion, and ethnic/gender diversity to the development of cost-effective force reduction programs. The model must be capable of producing multiple-year forecasts of the NSA's civilian force structure with a minimum fidelity level that allows for break-out by years-of-service, career field, and grade. Historical data will be utilized to estimate logistic regression parameters that will compose the internal workings of the model. The JMP statistical software package will be used for data analysis, while the model will be developed as a Microsoft Excel application

14. SUBJECT TERMS	15. NUMBER OF PAGES
Civilian Personnel Model, Logistic Regression, Principal Component Analysis, PCA, National Security Agency, NSA	102
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unlimited	Unlimited	Unlimited	UL