

NAVAL POSTGRADUATE SCHOOL Monterey, California



19970428 231

The Use of Survival Analysis in the Prediction of Attrition in Large Scale Personnel Flow Models

by

Robert R. Read

March 1997

Approved for public release; distribution is unlimited.

Prepared for: Deputy Chief of Staff, Personnel, U.S. Army
Washington, DC 20310-0300

DTIC QUALITY INSPECTED 1

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

Rear Admiral M. J. Evans
Superintendent

Richard Elster
Provost

This report was prepared for and funded by the Deputy Chief of Staff, Personnel,
U.S. Army, Washington, DC.

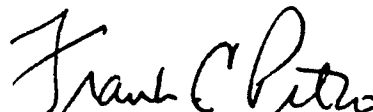
Reproduction of all or part of this report is authorized.

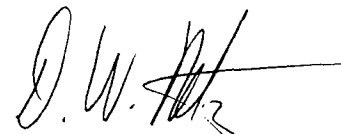
This report was prepared by:


ROBERT R. READ
Professor of Operations Research

Reviewed by:

Released by:


FRANK PETHO
Chairman
Department of Operations Research


DAVID W. NETZER
Associate Provost and Dean of Research

9/8/97

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE March 1997		3. REPORT TYPE AND DATES COVERED Technical
4. TITLE AND SUBTITLE The Use of Survival Analysis in the Prediction of Attrition in Large Scale Personnel Flow Models			5. FUNDING NUMBERS MIPR 6ESPMO 00029	
6. AUTHOR(S) Robert R. Read				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943			8. PERFORMING ORGANIZATION REPORT NUMBER NPS-OR-97-006	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Deputy Chief of Staff, Personnel, U.S. Army, Rm. 2C744 300 Army, Pentagon, Washington, DC 20310-0300			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES ..				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Survival analysis methods are applied to the problem of estimating attritions for the ELIM module of the Army's computerized personnel flow models. The use of the Cox proportional hazards model resulted in considerable computational overhead and the discrete proportional odds version (i.e., logit model) is examined instead. Since this model is supported by the generalized linear model software systems, some closely related models are considered and compared, specifically the linear, probit, and complementary log-log. The existing exponential smoothing method is also included. There are a number of important findings. First there is need to standardize the measures of performance. The present study uses average magnitude relative error for sets of forecast cells. Second, there is considerable variability in the forecastability of the attrition behavior of soldiers in the various personnel partitions; i.e., partitions made by cross classifying by education, mental group, contract term, etc. The logit, probit, and exponential smoothing models perform about equally well within the partitions when used without covariates and focusing on the involuntary losses. The inclusion of covariates in the generalized linear models appears to have a disturbing effect and the relative errors are larger than before. These errors have distinctive patterns, which at present, are unexplained.				
14. SUBJECT TERMS Attrition Rates, Survival Analysis, Logistic Regression			15. NUMBER OF PAGES 44	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

The Use of Survival Analysis in the Prediction of Attrition in Large Scale Personnel Flow Models

R.R. READ

Abstract

Survival analysis methods are applied to the problem of estimating attritions for the ELIM module of the Army's computerized personnel flow models. The use of the Cox proportional hazards model resulted in considerable computational overhead and the discrete proportional odds version (i.e., logit model) is examined instead. Since this model is supported by the generalized linear model software systems, some closely related models are considered and compared, specifically the linear, probit, and complementary log-log. The existing exponential smoothing method is also included.

There are a number of important findings. First there is need to standardize the measures of performance. The present study uses average magnitude relative error for sets of forecast cells. Second, there is considerable variability in the forecastability of the attrition behavior of soldiers in the various personnel partitions; i.e., partitions made by cross classifying by education, mental group, contract term, etc. The logit, probit, and exponential smoothing models perform about equally well within the partitions when used without covariates and focusing on the involuntary losses. The inclusion of covariates in the generalized linear models appears to have a disturbing effect and the relative errors are larger than before. These errors have distinctive patterns, which at present, are unexplained.

Acknowledgments

The research had the partial support the U.S. Army ODCSPER, under MIPR 6ESPMO 00029. The data cartridges were prepared from Army personnel files by GRC under the direction of Dr. S. Wilcox. The translation of the information for local use was made by Mike Whitaker. The author appreciates the assistance of Dr. S.E. Buttrey in a number of SPLUS programming issues.

1. INTRODUCTION

There is a general review taking place of the U.S. Army's computerized personnel management system. This review includes the examination of the forecasting process used to estimate numerous attrition rates. The present work examines the use of survival analysis to forecast attrition.

Personnel flow models in large organizations have numerous uses. These uses impact upon the structure chosen for data organization and the modeling process. Background of this type can be found in [GRC; 1985, 1989]. For the present purposes we cite the requirement that forecasts are required monthly for up to seven years into the future. We are concerned only with first term enlistees, i.e., non prior service personnel (nps). There is further partitioning according to cohort, C-group, and in the GRC work, loss type.

A cohort consists of all soldiers that enter the service in the same year and month. The C-group classification is based upon education, mental group, and sex. The details appear in Appendix A, and of course elsewhere [GRC; 1985, 1989]. Enlistees can commit themselves for periods ranging from two to six years, but over 98% of the contract terms are for either three or four years [GRC; 1985]. Since 1984, there has been provision to attach a variable amount (2 to 13 months) of training time to the contract term, and these are marked with a 'V' in the files. There are seven loss types utilized in the QFMD module [GRC; 1985]. That system is capable of making forecasts for up to 22,000 'projection cells'. It also partitions by term of enlistment. This module is designed to blend with other aspects of the global problem and treats extensions and reenlistments as losses.

The present work deals solely with C-groups 1, 2 and 4, and partitions them according to three and four year enlistment contracts. Further, we do not partition according to loss type. Losses are taken as leaving the Army for any reason, and do not include extensions or reenlistments. There is greater difficulty in forecasting near the ETS, or expiration of term of service date, largely because of the variable enlistment situation, the presence of sundry "early out" policies, and the fact that such losses are voluntary – that is the choice is in the hands of the soldier. Losses at earlier times are viewed as non voluntary, either by choice of the Army or by special circumstances.

In addition to the General Research modeling information there are also some views expressed by the Army Research Institute and their support contractors [Greenston, et al.]. The inclusion of

econometric factors in the redesign effort is new and provides an opportunity to include further information to the generation of loss rates and other aspects. For example, ACOL (annualized cost of leaving) modeling can be included and might be especially useful in the decision period prior to ETS. In the present work we choose to illustrate the manner in which covariates of general types can be included in the loss rate forecasts. Generally social and economic influences have been used, e.g. race, age, unemployment, etc. In fact, some of the partitioning variables (education, sex, etc.) could have been treated as covariates, but no such exploratory work has taken place. Many of the existing partitions have been made for administrative reasons and there may be no economy or incentive for including them as covariates.

A recent Master's Thesis [DeWald, 1996] contains some exploratory work with the use of Time Series methods for forecasting including Autoregressive Moving Averages (ARMA) and seasonal versions of exponential smoothing. It treats C-group 1 only.

Following this introduction Section 2 contains a brief description of survival analysis in terms most germane to the goal at hand. Section 3 contains a description of the data structure that must be dealt with. The material in Section 4 compares a variety of forecasting methods in the case of forecasting attritions without the use of covariates. The methods are extended to include covariates in Section 5, which also contains some speculations about what might be done to better the situation. The results are summarized in Section 6.

There are a number of appendices that contain supporting information and peripheral details. Specifically:

- A. C-group definitions and some accession distribution information.
- B. Formats of the raw data files and their conversions.
- C. Details of the exponential smoothing algorithm.
- D. Counting formulae for subsets of the data cell template of Section 2.
- E. Identification of the SPLUS program suites developed for the project.
- F. General formulas describing the GLIM method and its applications.

Another use of the report is to serve as a working paper for various derivative studies.

2. SURVIVAL ANALYSIS

Survival analysis is a body of statistical methods for studying the occurrence and timing of events. Although originally designed for the study of deaths, the methods have found considerable

use in many different kinds of social and natural sciences, including the onset of disease, earthquakes, recidivism, equipment failures, accidents, market crashes, job terminations, divorces and so on. These methods have been adapted, and often reinvented by workers in diverse fields. The names event history analysis, duration analysis, failure time analysis are commonly used but do not imply any substantial differences. However there can be a large variety of detail.

The basic data structure is the waiting time for an event to occur and a number of subjects or items are watched, waiting for that event for each. The event must be sharply defined relative to the time interval of inspection. For example the event of puberty or the onset of some diseases are events requiring duration and cannot be clearly identified as having occurred in a short time span. On the other hand things such as equipment failures, job terminations and marriages are clearly identified in time. The number of subjects under observation, the time period of inspection, and the recording of germane covariate information are details that affect the choice of method.

The event under observation in the present study is the event of a soldier leaving the Army, for any reason. There are many soldiers under observation; the inspection period is monthly, and the personnel files contain many potentially useful covariate items. For example, age, education, race, sex, and grade or job assignments may have an effect upon the rate of leaving. Also some general economic situations such as the civilian wage and unemployment situation may be effective during certain decision windows.

Basic Equations

Survival analysis can be developed either continuously in time or with a discrete set of time periods. The present application deals with monthly data, so we will present the basic ideas in terms of a set of discrete time periods measured from a known starting point. Let T be the random variable representing the number of periods that an individual survives (remains in the Army). The survivor function for that individual is the probability that he remains in the Army for at least t periods:

$$S(t) = \Pr\{T \geq t\}. \quad (1)$$

The hazard function (the attrition rate) can be expressed in terms of the survivor function as follows:

$$h(t) = [S(t-1) - S(t)]/S(t-1) \quad (2)$$

and is interpreted as the conditional probability that the individual attrites in the t^{th} period given he had survived $t - 1$ periods in the Army. There is a one to one relationship between the hazard function and the survivor function. The survivor function can be recovered from the hazard function using the product formula

$$S(t) = (1 - h(1))(1 - h(2)) \dots (1 - h(t)). \quad (3)$$

The number of personnel available at the beginning of period t is $N(t)$ and each individual is viewed as under the influence of the same survivor function. These subjects are viewed as independent trials, each with the same probability of attrition, $h(t)$. The expected number of losses in period t is $N(t)h(t)$. On the other hand, if one wants the expected number of losses during the $(t + k)^{\text{th}}$ period, that is k more periods into the future, and the most recent personnel inventory count is $N(t)$, then that expected number is

$$N(t) (1 - h(t)) (1 - h(t + 1)) \dots (1 - h(t + k - 1)) h(t + k), \quad (4)$$

that is, the expected number that survive the next $k - 1$ periods times the attrition rate for the subsequent period.

The modeling of survival functions may be in either the parametric or the non parametric mode. The parametric is useful for small populations, but one is more comfortable using the non parametric when the populations are large. Then the survivor and hazard functions are estimated directly from data. Specifically

$$\hat{S}(t) = N(t+1)/N(1) \quad (5)$$

and

$$\hat{h}(t) = (N(t) - N(t+1))/N(t). \quad (6)$$

The latter is the popular Kaplan-Meier estimator, [Cox & Oakes].

It is useful to record asymptotic distribution information. When the standard large sample theory for maximum likelihood estimators applies, [Cox & Oakes, p 50], then the expression

$$N(t)^{1/2} (\hat{h}(t) - h(t)) \quad (7)$$

will be asymptotically normal and independent (over the periods) with means equal to zero and variances estimated by

$$\hat{h}(t) [1 - \hat{h}(t)] / N(t). \quad (8)$$

From this follows the Greenwood formula for estimating the survivor function variance

$$\text{Var}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{j=1}^t \hat{h}(j) / \hat{S}(j) [1 - \hat{h}(j)]. \quad (9)$$

Some selected empirical survivor functions are shown in the Figure 2.1. The three and four year contract term curves are shown superimposed in order to emphasize the effect of differing terms and support the recommendation that the *C*-groups be partitioned by term of contract. A special interpretation must be made for these curves. Extensions and reenlistments were not treated as losses, but simply ignored. The curves go well beyond the 36 and 48 month terms indicated. This is because the training time is added to the contract and is a variable number of months.

Turning to the inclusion of covariates in the forecasting of attritions, there is a difficulty in applying the popular Cox proportional hazards model, due largely to the large number of tied observations. Our application deals with large numbers of individuals, and the recording of accessions and attritions is accumulated by month in the small tracking file. It seems wise to treat the data as discrete and use the proportional odds model instead, [Cox & Oakes, pp 101 ff]. That is, there is a baseline ratio of the hazard rate to the continuation rate and the odds ratio for individuals with covariate structure z is proportional to the baseline ratio, and the proportionality Ψ is a function of z , and the set of parameters β that leverage their importance. This is the discrete logistic model. That is

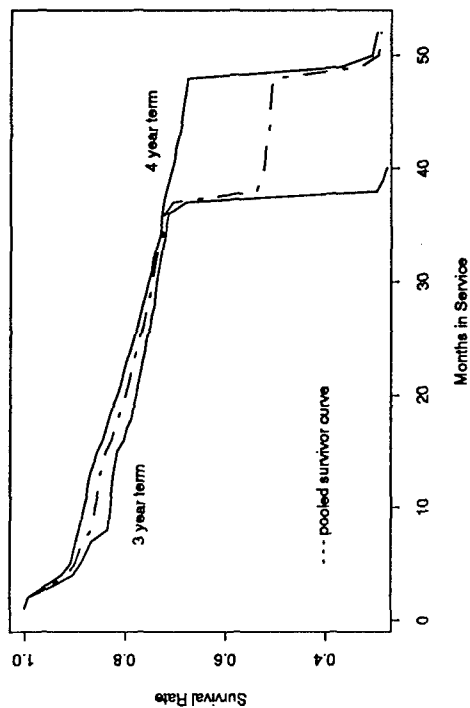
$$\frac{h_z(s)}{1 - h_z(s)} = \Psi(z; \beta) \frac{h_0(s)}{1 - h_0(s)}. \quad (10)$$

The crossing of the survivor functions in the Figure 2.1 is not consistent with the proportional odds modeling, so further support is given to the partitioning of personnel contract term.

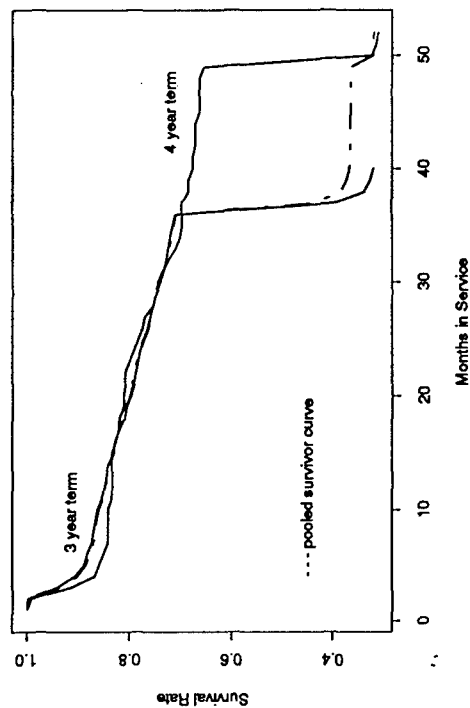
General Considerations

The concept of partitioning into groups can be taken too far. Groups that are too small can exhibit statistical instability. This certainly occurs in applications to other fields (e.g. surviving a particular form of surgery) in which the group sizes are not large inherently. In these instances, additional sources of variability may be accounted for using measured covariates such as age, ethnic background, health history, etc. Some such sets of covariates may appear as a further partitioning of the group, but statistical regression type modeling can be applied and will prevent

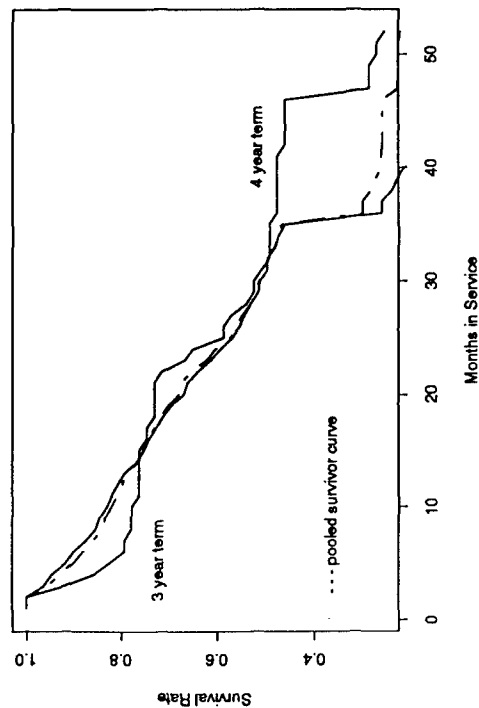
Cohort 198302; C group 1



Cohort 198403; C group 2



Cohort 198409; C group 4



Cohort 198504; C group 4

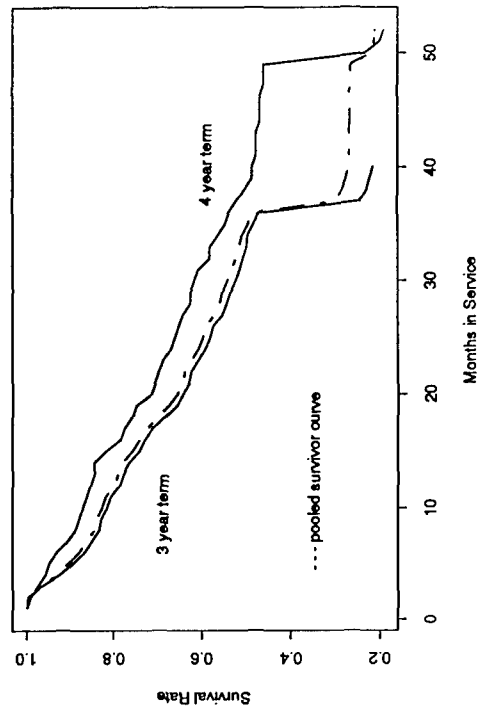


Figure 2.1 Selected Survivor Curves Pairing the 3 and 4 Year Contracts

the dilution of the data base. One can develop a baseline hazard function and some modification rules for adjusting the hazard to suit the particular combination of covariates. Other types of covariates may be time dependent, affecting different individuals at differing times in the course of the study. These too can be accommodated using regression type methods.

Once a particular group or class of soldiers has been identified, then we have the further consideration that such a group of recruits will be accessed into the Army once every month. Such a monthly collection of recruits identifies a cohort. This presents a new issue. The inauguration times of the cohorts are staggered and one must face the issue of whether each cohort should be treated as having its own survival function, or can they be pooled into a common population. It seems unwise to pool a large number of them over a large number of starting periods, but there is evidence that they can be safely pooled over a small number. For now, the default position is that each cohort will be treated as having its own survival function. (Some combining will take place during the study, however.)

The goal of our study is to examine whether survival analysis can be successfully applied as the attrition rate generator in the Army's personnel flow planning system. The usage is to forecast attritions for many months into the future based upon the historical data of the past. We will be testing some methods for doing this. The efficacy will be measured by cross validation. That is, some historical data will be used for the making of forecasts and the effect will be measured by how well they forecast other data. That is, the test is made on data different from that used to generate the forecast.

Measure of Effectiveness

A measure is needed to compare the performances of the several methods tested. Relative error of forecast is used for each cell, where the cells are defined in terms of the data template of the next section. Basically, for a given partition designation, a cell is a month into the future and a cohort.

Let pred be the predicted number of survivors of a cell and let act be the actual number that survived. Then compute

$$\text{Relative error} = (\text{pred} - \text{act}) / \text{act} \quad (11)$$

and this is the relative error in the predicted number of survivors. Notice that since the error in the number of losses forecast is the negative of the numerator, this value (11) is the negative of the

relative error of predicted losses. Generally, the magnitudes of the relative errors can be expected to increase as the cells are taken deeper into the future. The growth of this figure with respect to time is of interest as are some other aspects as well. Generally the average magnitudes of these errors are recorded in the tables of Section 4 for various time blocks. The graphical displays provide relative error information in a sequential order to be described.

The above measure is different from that used in [DeWald, p 29]. There, DeWald was interested in forecasting strength at a particular time. He aggregated the cell forecasts and actuals over the mis (months in service) variable while holding time fixed. Then he computed the relative errors. For purposes of summary he averaged them over blocks of time, allowing positive and negative input to offset. Thus the methods treated there cannot be compared directly with those used here.

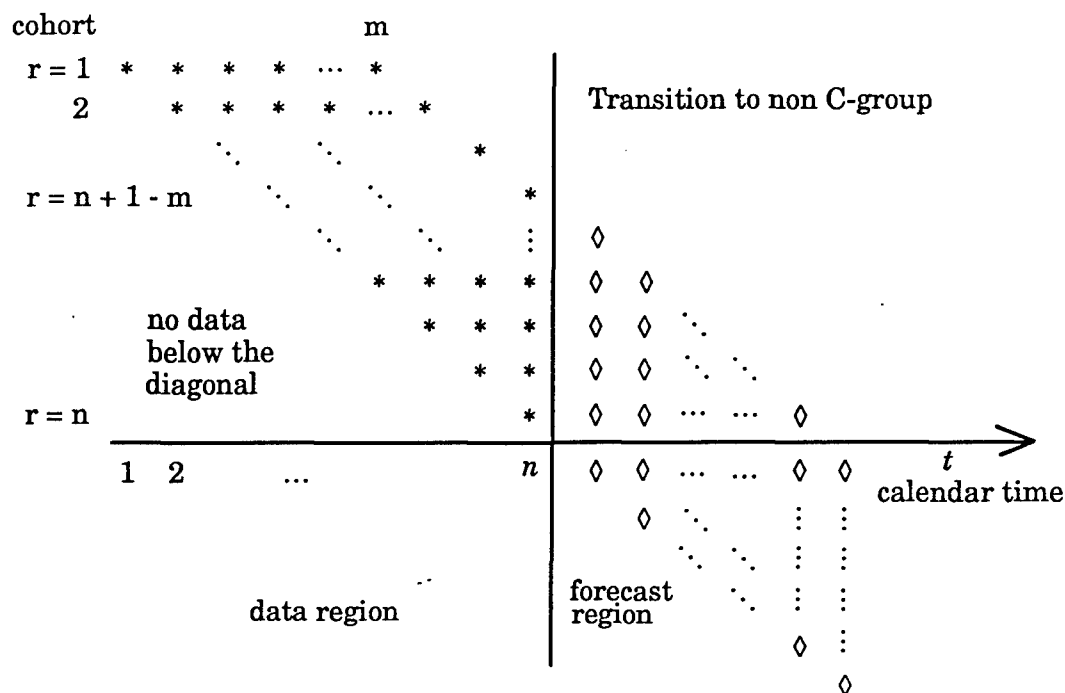
The author is unaware of the performance of the system presently in use. But any such comparison should be made using a common measure. [Dillaber] alludes to an "average error on loss projections" of about 5% and an "average error on man-year projections" of about 0.1%. GRC states [GRC; 1989] "strength projections ... have achieved an accuracy of within +/-0.5% for at least a 12-month horizon. We know not how these values were computed.

3. DATA STRUCTURE

Envision the following trapezoidal template for a given *C*-group/contract term. Each cell ($*$ or $\diamond$) refers to a cohort ($r = 1, \dots, m + n$) and a time period ($t = 1, \dots, m + n$). Time here is ordinal and indexes months. The most recently completed month is $t = n$.

The vertical line separates the past from the future. The horizontal line partitions the forecast region so that: below the line we forecast losses for cohorts that have either not yet convened or not completed the first period. Above the line we must forecast losses for existing cohorts that have completed at least one period. The quantity m serves as the cohort length parameter. In the present work it serves to separate the involuntary losses from the commingling of the voluntary and involuntary losses. Beyond m periods is a time window that has a more variable attrition structure. After ETS the soldiers are placed in a non *C*-group status. In this way the *C*-group does not become too thinly populated. We view m as a tuning constant that must be carefully selected. We are not yet working with either the commingling or the the non *C*-group.

Data Cell Template



There is also need to measure time relative to the start of a cohort; i.e., month in service (mis); the index s will be used, and $s = 1, \dots, m$.

The periods are months and the parameter m is related to the contract length of the C-group. For the present we are focusing on involuntary losses and choose m to be three months fewer than the contract length. Thus, if the contract length is four years then m is chosen to be 45. The parameter n is equal to 72 for six years of data, given the number of periods used in the data base. It is always assumed that $n > m$.

The data in hand must be divided into two sets, one for forecasting and one for testing. The former must precede the latter in time. The data template helps us to visualize the process and serves as a basis for discussion of some important details. Calendar time is indexed with the symbol t , which is the ordinal index of the periods. Our monthly data begins in 198301 (1 January, 1983). This is $t = 1$, and at the n^{th} month the estimation interval stops. For our work this is $n = 72$, or 198812. These data cells are marked with asterisks. They include many cohorts, some

in their entirety and some partially. Forecasts are required for all cells with t larger than n . These are marked with diamonds. Thus the data for 198901 and beyond are reserved for validation. (Of course in an operational context, the value $t = n$ would index the most recently completed month.)

The forecast region is to the right of the vertical axis and is marked with diamonds. Those above the horizontal axis joined with the first row below that axis form the cells referred to as the critical triangle. This triangle has the "most meat". In each row the initial inventory counts will be known (first column of diamonds to the right of the vertical axis), and the values of the covariates will be known, i.e., the racial mix and the age group at enlistment structure are being used in this study. No such information will be available for cells below the critical triangle. The forecasting in this latter region will be cruder – conceptually easy but not as accurate. Here the forecasts can only be made in a relative sense. The hazard rates can be used, but there are no inventory numbers to apply them to. One could use planned accessions, but no such information is available for this study. One can only expect the error of forecast rate to be higher. This set of cells will be referred to as those below the critical triangle.

It is appropriate to comment on the cells above the critical triangle. Although not marked they represent cells for which forecasts are needed. Those cells near $t = n$ represent soldiers near the end of their first contract. Some will have extended, or reenlisted, or simply left the Army by allowing the commitment to terminate, i.e., the voluntary attritions.. Those cells farther from $t = n$ represent soldiers who have made a fresh commitment and are once again at risk to those effects that come under the umbrella of involuntary attritions. Since the present study is concerned only with non prior service, first term enlistees, the boundary that cuts across these cells is vague.

The policy has been to drop the C-group designation at the end of the first term of service. Thus all such soldiers become pooled into a "non C-group" upon continuing beyond the first contract. Presumably they maintain their cohort identity.

Let us return the discussion to those cells in the region in which voluntary and non voluntary losses are commingled. Some exploratory data analysis in this area has revealed considerable variability. This is a region in which some careful modeling can be applied. Some ACOL models (annualized cost of leaving) have been proposed [Greenston, et al.; Smith, et al.] and of course they should be tested. A boundary parameter, m , has been introduced because of the special nature of this region. It represents the maximum length of service time in a cohort before the voluntary loss effect become important. The idea is to separate the less variable non voluntary loss environment from the more volatile environment in which the two types are commingled.

This study concentrates on the forecasting of attritions in the critical triangle cells. The forecasts must be computed from the information in the cells marked with asterisks prior to $t = n$. There is choice in how to use this information.

In parametric applications of survival analysis, such as reliability of equipment, historical failure data are used to fit a parametric model to the time between failures data. For example the Weibull distribution is a popular choice. Such processes are time homogeneous, i.e., it matters not where the origin of observation is located, and once the model is fitted it can be used to forecast the next failure time. Such forecasts are used to decide upon maintenance and repair policies. In the present application an attrition plays the role of a failure, and each cohort could be used by itself for the historical data base. It seems that the assumption of time homogeneity is less supportable, largely because we are dealing with a human resource. A time homogeneity assumption would allow us to fit a survival curve to each cohort and pay no attention to its inauguration date. The staggered starts wouldn't matter. Although this might be useful for a few cohorts whose starting times are close together, it seems unwise to do this broadly. This has been mentioned earlier. On the other hand this option might be held in reserve for study using the smaller C -groups. In these cases the small personnel numbers may be inadequate for the use of non parametric survival models.

There is another practical aspect to this that should be mentioned. Looking at the data cell template, we see that some cohorts have much historical data. For these it may be possible to build an extrapolation curve that could be used for forecasting. In these cases the forecast cells (diamonds) will be few in number. On the other hand when we look at those cohorts for which there are many forecast cells (diamonds), we see that the number of cells on which to base a forecast is diminished, and goes to zero in the extreme. Accordingly, this particular approach to using the historical data has an awkward aspect. Instead of forecasting 'line by line' along the cohorts of the data template, perhaps more of the estimation cells should be used, thereby cross cutting the various cohorts. This indeed is what has been done in the past [GRC; 1985, 1989].

Indexing Conventions

The data cell template in the diagram is layed out in the (t, r) plane. That is, the horizontal index is t and it refers to an ordinal number of time periods relative to the calendar, or real time. The vertical index is r and it refers to the ordinal number of the cohort, i.e., the period during which the cohort begins. A third symbol, s , marks the number of periods in service measured from the start of the cohort. Now the data format in Appendix B in Table B.2, is in the (s, r) plane.

This point should be noted because the first column is in a 198xxx notational format, e.g. 198301 refers to January of 1983. In the programs, this time format must be used and can refer either to real time or to the cohort starting date. For use of the input data, it refers to the latter; the convening date of the cohort. The following few columns mark the covariate combinations, and after that the headings are $s = 1, 2, 3, \dots$, for months in service. The data entries are integers representing the at risk counts, or personnel inventories at the beginning of the period.

Data visualization is important and it is useful to create a mental image of the two forms described above. Cells identified in the (t, r) plane are easily found in the (s, r) plane simply by using the relation $s = t + 1 - r$. In the (t, r) plane the 'wall' separating the estimation cells from the validation cells is at the end of the $t = n$ period and is visualized as a vertical line. However this line, in the (s, r) plane, is a diagonal one whose equation is $r = n + 1 - s$.

The prediction scheme that has been in use [GRC; 1985] is described in the (s, r) plane and utilizes a rectangular set of cells there. In order to visualize this in terms of our cell template in the (t, r) plane, one must extend the line $r = t + 1 - s$ to the "northwest" from the point $(n, 1)$ until it crosses the $t = 1$ axis. In this way there are a full set of n cells on each diagonal associated with each constant value of s . Their exponential smoothing algorithm is applied for each fixed value of s and based upon a common number of cells. Attention is drawn to the way that this rule cross cuts the cohorts previous to $r = 1$.

In the present work, when cells having constant month in service, s , are used for estimation, we limit ourselves to those cells actually in the data template. Our data derived from the small tracking files is in this form, and the more extensive information is not readily available. The length of the cross cut data string is $n + 1 - s$ for each s .

A number of estimation schemes are applied in this fashion. The comparison of methods is made both without and with the use of covariates.

4. FORECASTING ATTRITIONS WITHOUT THE USE OF COVARIATES

The fundamental prediction scheme can be applied using any number of averages to generate the consensus hazard values to be applied to the future. Such is done in this section and the performances are compared. Also some graphical work is presented and the viewer can observe the relative errors of prediction grow as one moves further into the future.

We can compute a Kaplan-Meier estimate of the hazard for every cell in the estimation region of the data template, i.e., those marked with asterisks. First use the notation of the (s, r) plane. Let $N(s, r)$ be the beginning cell inventory (number at risk) in cell (s, r) for mis = s and cohort = r . The estimated hazard is

$$\hat{h}(s, r) = N\{(s, r) - N(s+1, r)\} / N(s, r). \quad (12)$$

For each fixed value of s there are $n+1-s$ values of r . We will average these in various ways, each type producing a consensus value, $\hat{h}(s)$, which will be used to predict the survivors for each cell in the critical triangle. But before describing the various averaging methods, let us show how the $\hat{h}(s)$ values will be used this prediction process.

The earlier equation (4) will be used, but the double subscripting must be faced. Now it is convenient to change to the notation of the (t, r) plane. Let $NN(n+1, r)$ be the (known) number of personnel at risk in the diamond cells immediately to the right of the 'wall'. That is, $t = n+1$ with the r values in increasing order starting at the top of the critical triangle, $r = n+2-m, n+1-m, \dots, n+1$. (The number of diamonds in a column of the data template is m .) The predicted number of survivors in any diamond cell is the appropriate 'at risk' value when crossing the 'wall' times the product of the proper number of survivor rates. Let $SS(t, r)$ be the forecast number of survivors in cell (t, r) . Then, using the consensus values, $\hat{h}(s)$,

$$SS(t, r) = NN(n+1, r) (1 - \hat{h}(n+2-r)) (1 - \hat{h}(n+3-r)) \dots (1 - \hat{h}(t+1-r)) \quad (13)$$

for $t = n+1, n+2, \dots, n+m$ and $r = t+1-m, t+2-m, \dots, n+1$.

Returning to the question of estimating the consensus hazard values, $\hat{h}(s)$, from the Kaplan-Meier data, $\{\hat{h}(s, r)\}$, we consider six estimators each of which results from an averaging process over the $n+1-s$ values available when s is held fixed. All have supporting rationale.

- a. Direct linear average. This is merely the arithmetic mean of the specified values.
- b. Weighted linear average. This is the weighted arithmetic average, the weights being the inverse variances of the hazard estimates, eq.(8).
- c. Logit transform. This time we average the log-odds and invert the transform. This method appears in survival studies and conforms with the log odds model (10) without covariates. Briefly, the values

$$y(s) = \text{ave}_r \left\{ \log \left(\hat{h}(s, r) / (1 - \hat{h}(s, r)) \right) \right\} \text{ are inverted to } \hat{h}(s) = 1 / (1 + \exp(-y(s))). \quad (14)$$

- d. Probit transform. The inverse Gaussian distribution is applied to the hazard values, the results are averaged and then inverted using the Gaussian distribution, [Allison].
- e. Complementary log-log transform. Also used in survival analysis, [Allison]. It is another average transform method. Specifically

$$y(s) = \text{ave}_r \left\{ \log \left(-\log(1 - \hat{h}(s, r)) \right) \right\} \text{ and } \hat{h}(s) = 1 - \exp(-\exp(y(s))). \quad (15)$$

- f. Exponential smoothing. This familiar technique is a weighted average, but the weights diminish in size exponentially as the terms go further into the past. The version that we used is described in Appendix C. The smoothing constant used in the comparisons below is $\alpha = 0.05$.

Table 4.1 summarizes the computations applied to the partitions we have used. The entries are the average magnitudes of the relative errors for the six methods aggregated by year into the future. The groups are marked with their C-group number followed by a decimal point and then the term of enlistment. For example C2.4 refers to C-group 2 and the four year contract term. The logit, probit, and exponential smoothing methods appear to perform a bit better than the others. However, Table 4.1 reveals that there are clear distinctions among the partitions; some are more difficult to predict than others. The number of cells in the first year set of forecasts is 330 for three year term groups, and 474 for four year term groups. The number of cells for the second, third and fourth years into the future appear in Table 4.1 and are marked "size".

Table 4.1. Average Magnitude Relative Error by Year into the Future and by Method

Data = C1.3				data = C1.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330.	186.	45.	size	474.	330.	186.	45.
lin.ave	0.01666	0.04996	0.07830	lin.ave	0.01188	0.01419	0.01951	0.2372
lin.wtd	0.01671	0.04893	0.07592	lin.wtd	0.01210	0.01429	0.01941	0.2341
logit	0.01675	0.04817	0.07404	logit	0.01227	0.01442	0.01944	0.2317
probit	0.01759	0.04383	0.05959	probit	0.01409	0.01695	0.02168	0.2112
logmlog	0.01792	0.06173	0.09959	logmlog	0.01016	0.01620	0.02731	0.2726
expsm	0.01869	0.04387	0.05966	expsm	0.01481	0.01858	0.02267	0.2068

data = C2.3				data = C2.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330.	186.	45.	size	474.	330.	186.	45.
lin.ave	0.01438	0.03095	0.05800	lin.ave	0.01063	0.03384	0.05804	0.2747
lin.wtd	0.01445	0.03040	0.05667	lin.wtd	0.01047	0.02992	0.05044	0.2623
logit	0.01452	0.03001	0.05594	logit	0.01049	0.02712	0.04470	0.2529
probit	0.01557	0.02846	0.05102	probit	0.01433	0.01721	0.02192	0.1781
logmlog	0.01425	0.03679	0.07444	logmlog	0.01604	0.06568	0.11410	0.3671
expsm	0.01806	0.03169	0.05028	expsm	0.01540	0.01851	0.02504	0.1705

data = C4.3				data = C4.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330.	186.	45.	size	474.	330.	186.	45.
lin.ave	0.05384	0.1490	0.17583	lin.ave	0.03701	0.12090	0.1674	0.6479
lin.wtd	0.05482	0.1502	0.1808	lin.wtd	0.03461	0.10270	0.1375	0.5848
logit	0.05562	0.1513	0.1845	logit	0.03330	0.09032	0.1187	0.5350
probit	0.06211	0.1614	0.2081	probit	0.03916	0.06506	0.1030	0.2300
logmlog	0.05048	0.1421	0.1508	logmlog	0.06001	0.23350	0.3654	1.0300
expsm	0.06326	0.1643	0.2086	expsm	0.04102	0.07005	0.1135	0.2233

Since the exponential smoothing methods can change when the smoothing constant is changed we should take a brief look at the performance of exponential smoothing for several values of the smoothing constant. The value used above is $\alpha = .05$. Let us compare the effect when $\alpha = .01$ and $\alpha = .1$ as well. Table 4.2 shows the the situation is quite stable and there is little difference in performance for these three values, although there may be a deteriorating trend as α increases.

Table 4.2. Average Magnitude Relative Error for the Exponential Smooth Method Using Three Values of the Constant

Data = C1.3				data = C1.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330	186	45	size	474	330	186	45
expsm.01	0.01785	0.04386	0.05968	expsm.01	0.01423	0.01730	0.02183	0.2105
expsm.05	0.01870	0.04387	0.05966	expsm.05	0.01481	0.01858	0.02267	0.2068
expsm.1	0.01918	0.04381	0.05936	expsm.1	0.01526	0.01947	0.02384	0.2020
data = C2.3				data = C2.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330	186	45	size	474	330	186	45
expsm.01	0.01604	0.02899	0.05075	expsm.01	0.01449	0.01739	0.02239	0.1771
expsm.05	0.01806	0.03169	0.05028	expsm.05	0.01540	0.01851	0.02504	0.1705
expsm.1	0.01987	0.03487	0.05093	expsm.1	0.01650	0.01984	0.02785	0.1610
data = C4.3				data = C4.4				
	1st yr	2nd yr	3rd yr		1st yr	2nd yr	3rd yr	4th yr
size	330	186	45	size	474	330	186	45
expsm.01	0.06235	0.1621	0.2082	expsm.01	0.03960	0.06628	0.1056	0.2283
expsm.05	0.06326	0.1643	0.2086	expsm.05	0.04102	0.07005	0.1135	0.2233
expsm.1	0.06420	0.1663	0.2094	expsm.1	0.04177	0.07161	0.1171	0.2221

This opportunity is taken to display some quantitative information on the role of partitioning by term of enlistment. We have computed the average magnitude relative error rates for C-group data that has not been separated by term of enlistment. Table 4.3 utilizes the logit method only. The data from C-groups 1, 2, 4 have been pooled. The three and four year enlistment term data have been aggregated into single 36 month sets for each cohort. (It was not convenient to include the fourth year because the derived "at risk" data set (Appendix B) did not carry the mis variable that far.)

Table 4.3. Average Magnitude Relative Error When Constant Terms are Merged

	1st yr	2nd yr	3rd yr
size	366	222	78
C-1	0.0237	0.0313	0.0445
C-2	0.0291	0.0408	0.1054
C-4	0.0582	0.1114	0.1949

A comparison of these values with those of parts of Table 4.1 that separate by term of contract indicate that the separation advantage is significant.

Returning to the comparison of methods, a graphic view of the relative errors provides more detail. There are too many cases for an extensive view, but some have been selected for illustration, and appear in Figure 4.1.

Let us establish the order of weaving our way through the prediction cells in the critical triangle. The first cell treated is the one at the upper left corner. For three year term data this cell has month-in-service value $s = 33$. From here we decrement s downward through the triangle until $s = 1$. This completes all forecasts that extend exactly one month into the future. Then we return to the top of the triangle and treat the column of cells for exactly two months into the future, with s starting again at 33 and this time decrementing to $s = 2$. The pattern is continued in this "vertical strip" system until all cells in the triangle are treated.

This is the order of plotting in the Figure 4.1. There are four graphs each with a marked case. The method is different for each, but we can see from Table 4.1 that the forecasting method contributes little to the information in the graphs. The important sources of variability are the C-groups by contract term. In particular, C-group 4 with a three year term is quite noisy regardless of the estimation method applied. The two four year term cases both have an upturn at about the same distance into the future. This is a Desert Storm effect. The future time is 1992, a time when some aggressive 'early out' programs were initiated.

5. FORECASTING WITH COVARIATES

At the beginning of the research we had planned to use the Cox proportional hazards method in conjunction with SAS software (or other industrial strength programs). This resulted in some extremely large running times. Investigation into this problem led us to follow the recommendations of [Allison, ch 7], which contains much experiential information. We were led to the use of generalized linear models. The discrete proportional odds model, eq.(10) fits into this framework [Allison, p 212] as does some others; the probit and the complementary log-log, [Allison, pp 216 ff]. (This is the reason that these models were applied in Section 4.) Also at this earlier time we used unemployment as a covariate. Although it tested as significant, it did not enhance the predictability in a noticeable way. It has not been included in what follows.

To continue the research, we worked with the SPLUS system, which has both survival analysis and generalized linear model capabilities, is PC based, and has rather good graphics. The linear, logit, probit, and complementary log-log methods introduced in Section 4 are supported by it. These are included as options in our programs.

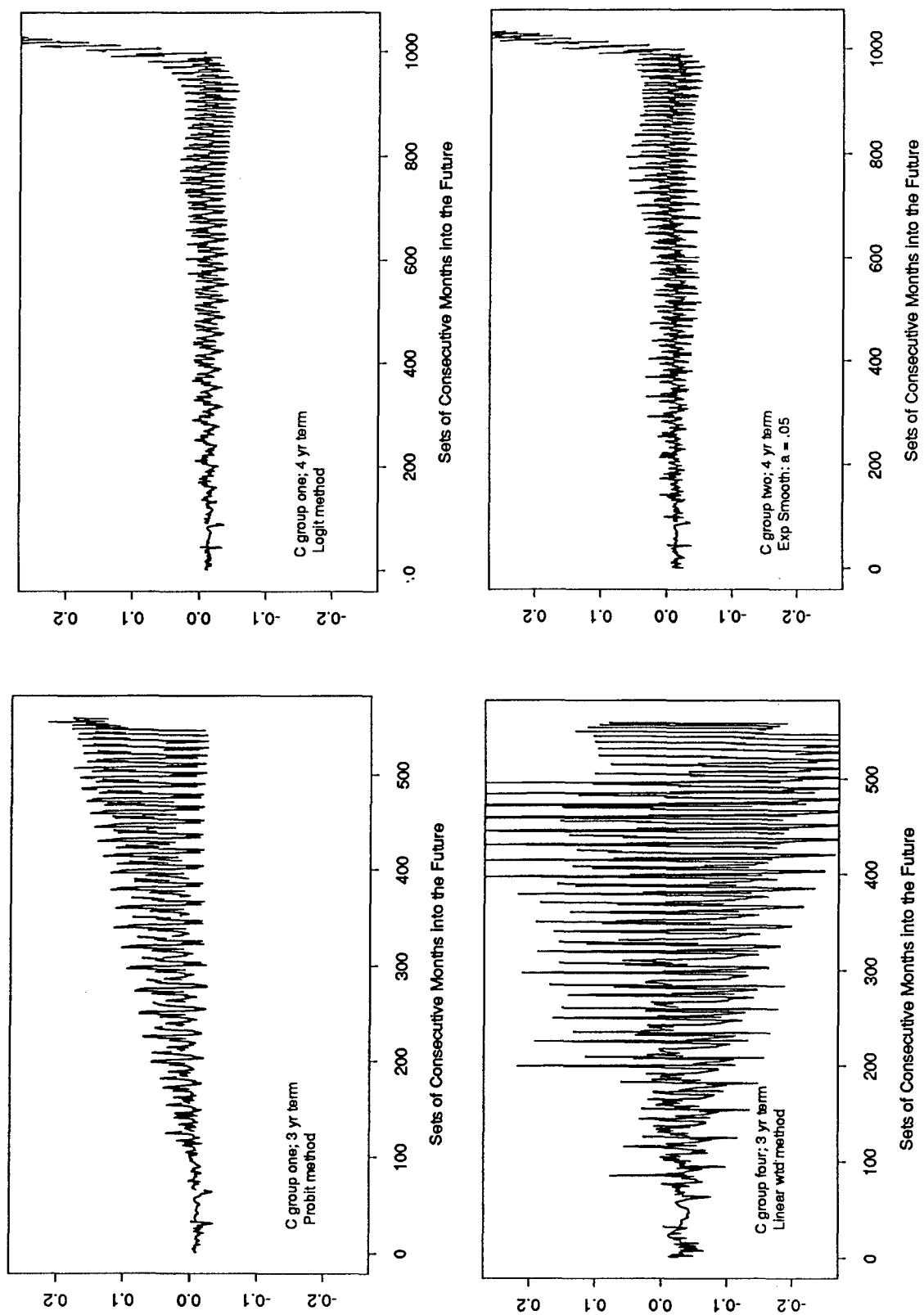


Figure 4.1. Selected Relative Error Curves for Cells in the Critical Triangle

Two covariates were selected for testing, age and race. Both are cohort related, that is, their effect modifies the survivor functions of a cohort. The cohort length problem mentioned earlier comes into play. If there is an adequate number of mis values for assessing the effect of the covariates, then there are few validation cells. Conversely, if there are many validation cells, then there is not much data to estimate the effects of the covariates.

The general structure of these models is described first. This is followed by some details of implementation and then the results.

The covariates selected are treated as factors; age has six levels and race has three. See Appendix B. For each cohort the derived input data provides up to eighteen strata of 'at risk' counts. These are the personnel inventories at the start of the period and from these we can compute m Kaplan-Meier hazard estimates for each strata. These values are transformed by the function of our choice; identity, logit, probit, or complementary log-log. Call these values y and apply the statistical model

$$y = X\beta + \varepsilon \quad (16)$$

where X is a design matrix of sixteen columns (two fewer than 18 so as to avoid singularity due to overparameterization). The number of rows is the smaller of m and $n + 1 - r$ where r is the index of the cohort; β is the parameter vector to be estimated; and ε is the error term. In general the error terms do not have constant variance; indeed the distribution involved is the binomial distribution. The generalized linear model function will provide us with maximum likelihood estimates for the model (16). It will also provide a set of 'fitted' values for the y 's and the variances. This will be called the GLIM step, (Generalized Linear Models).

At this point, we must choose the factor levels to serve as the baseline. We chose the second age level (over 18 but less than 20) and the first race level (white). The set of fitted y 's, augmented with the proper linear terms to reflect this combination, are subjected to the inverse transform and serve as baseline hazard values. (In the case of the identity transform it is possible to acquire some negative values. Any such are replaced with zeros.)

What has been described will work unless the cohort index, r , is more than $n - 15$. Then the number of rows of X will be fewer than the number of columns. Our response to this problem has been to combine contiguous cohorts into echelons which are modeled to have a common hazard function as well as common response to the covariates. Notice that this is a bit different from treating the several cohorts as strata with common response to the covariates. In this latter case

each stratum has its own hazard function. This structure also could be used, but there is virtually no difference in the hazard functions. We have chosen to use three contiguous cohorts to form an echelon. It will be seen that the fits are good.

Having made this choice, the three cohorts in each quarter year will share the coefficient estimates of the various factor levels.

Let us examine how the forecasting technique is affected by this. The number of survivors of period n serve as the inventory figures to forecast survivors at all periods beyond the 'wall'. Formerly there were m such counts required, but now there are $18m$; each cell must have 18 to allow for the covariate combinations. (Of course some of these counts may be zero.) In order to project survivors for the cells in the critical triangle we also require the consensus hazard values and these must be augmented to reflect the age-race combinations. The method of augmentation is a function of the transformation chosen for the GLIM step and need not be described further. The mathematics of this is outlined in Appendix F.

Two types of estimates were used for the consensus hazard values. The first is the weighted average method described in Section 4, and this is called 'grandm' for grand mean. The proper weights are part of the GLIM output. The second is called 'current'. It simply uses the m baseline hazard values available for $t = n$. That is, the most recent hazard value is chosen for each mis value s . These are the choices for type 1 input to the programs.

This done then the forecasts are made in the same way as before, but now there are eighteen times as many, one for each covariate combination. The eighteen forecasts for each cell in the critical triangle are added together to obtain a total for use in the relative error computation.

Some results are presented below. First, an analysis of deviance table, Table 5.1, is presented to illustrate that the echelon modeling is valid and that the covariate factors are significantly different from zero. The models fit the data well. The logit method and echelons 14 and 15 using the C1.3 data were selected. The two contiguous periods are April through June and July through September of 1986. Contiguous periods were chosen to indicate the variability with time of the factor level coefficients. It is rather large. But first let us look at the analysis of deviance.

It is presented in nested model form [Agresti, p 211 ff]. It begins with the null model, that is, the model that is expressed by a single term (a constant rate). Then the three factors are added sequentially in the order mis (month in service), race, and age.

Table 5.1. Analysis of Deviance Table - Logit Model

April-June 1986				
model	df	deviance	residual.df	residual.deviance
null			4619	5158.325
mis	29	1097.730	4590	4060.595
race	2	38.238	4588	4022.358
age	5	352.854	4583	3669.503
July-September 1986				
null			3902	3745.202
mis	26	542.6450	3876	3202.557
race	2	35.9586	3874	3166.599
age	5	82.4246	3869	3084.174

The residual deviance serves as a test statistic for the adequacy of the model that includes the factor in a line and all factors above that line. If the model is valid, then asymptotically its distribution is chi-squared with the residual degrees of freedom. Ad hoc evaluations can be made using the facts that a chi-squared random variable has a mean equal to its degrees of freedom and variance equal to twice its degrees of freedom. The columns marked deviance and df are the first differences (resp.) of the two outer columns. The same ad hoc evaluation rules apply to judge the significance of the individual factors.

The attritions are rather rare events and the models are being fitted to the tail of the logistics curve. The counts themselves are small relative to the inventory numbers. Because of this there is doubt as to the appropriateness of using the asymptotic chi square distribution theory to judge the overall quality of fit. On the other hand, the deviances actually become smaller than the residual degrees of freedom. Such an occurrence can be a fortiori evidence of an acceptable fit. We imagine that the application of exact distribution computation would support the models. Of course the null model in the first set is an exception, but the inclusion of the additional factors leads to acceptable models. Further, the attribute of over fitting is not objectionable when the goal is forecasting.

More importantly, the use of chi square tests is more palatable when judging the change in the deviance as the various factors are added to the model. In Table 5.1 above the differences in deviances are typically seven or more times the number of degrees of freedom. All of the factors are statistically significant. The fitting can only improve if there were fewer cohorts in an echelon.

Several dozen such tables have been examined. The mis and age factors are always significant. There are a number of cases in which the race factor is not significant.

Next let us examine how the estimates change from one period to the next. The values are in the linear scale of the logit transformation, making them difficult to interpret. But since they are fixed effects, it seems that the temporal changes in their values should be somewhat smooth. Such hopes have been disappointing. The values below illustrate. The SPLUS software uses the contrast system that sets the coefficient of the last level of a factor equal to zero. In our case those are race = other and age = over 28.

Table 5.2. Factor Level Coefficient Estimator for Two Contiguous Quarters

April - June 1986							
Intercept	race1	race2	age1	age2	age3	age4	age5
-4.821516	0.03166253	0.105502	0.05607	0.08760	0.15530	0.16165	0.05026
July - September 1986							
Intercept	race1	race2	age1	age2	age3	age4	age5
-4.795383	-0.05225	0.11016	-0.2738	-0.05478	0.00006	0.03373	0.03818

The levels of the race and age factors are identified in Appendix B.

Two graphs of the forecast relative errors have been selected and appear in Figure 5.1. Generally our graphs of this type are poor, certainly not as good as those generated without the use of covariates. The year into the future error curves are clearly separated, at least in the early going. The logit method has been chosen in conjunction with the grand mean averaging of the baseline values for each fixed value of s . (The use of the "current" option did not change things much.) The probit method works about as well. The other two methods (linear and complementary log-log) do not perform as well. Again the partition chosen has as much effect as it did in Section 4.

Attention is drawn to the following features of the two graphs.

- i. The relative errors are larger than those produced earlier without the use of covariates.
- ii. The curves show very distinct patterns. They are monotone beginning with $s = 33$ until s achieves its minimum for the year. After two or three months, some secondary patterns emerge which are also quite distinctive. One pattern is for the central values of s and some others for the extreme values of s . As of this writing we know not the causes of these patterns, but they are strong enough to be correctable.

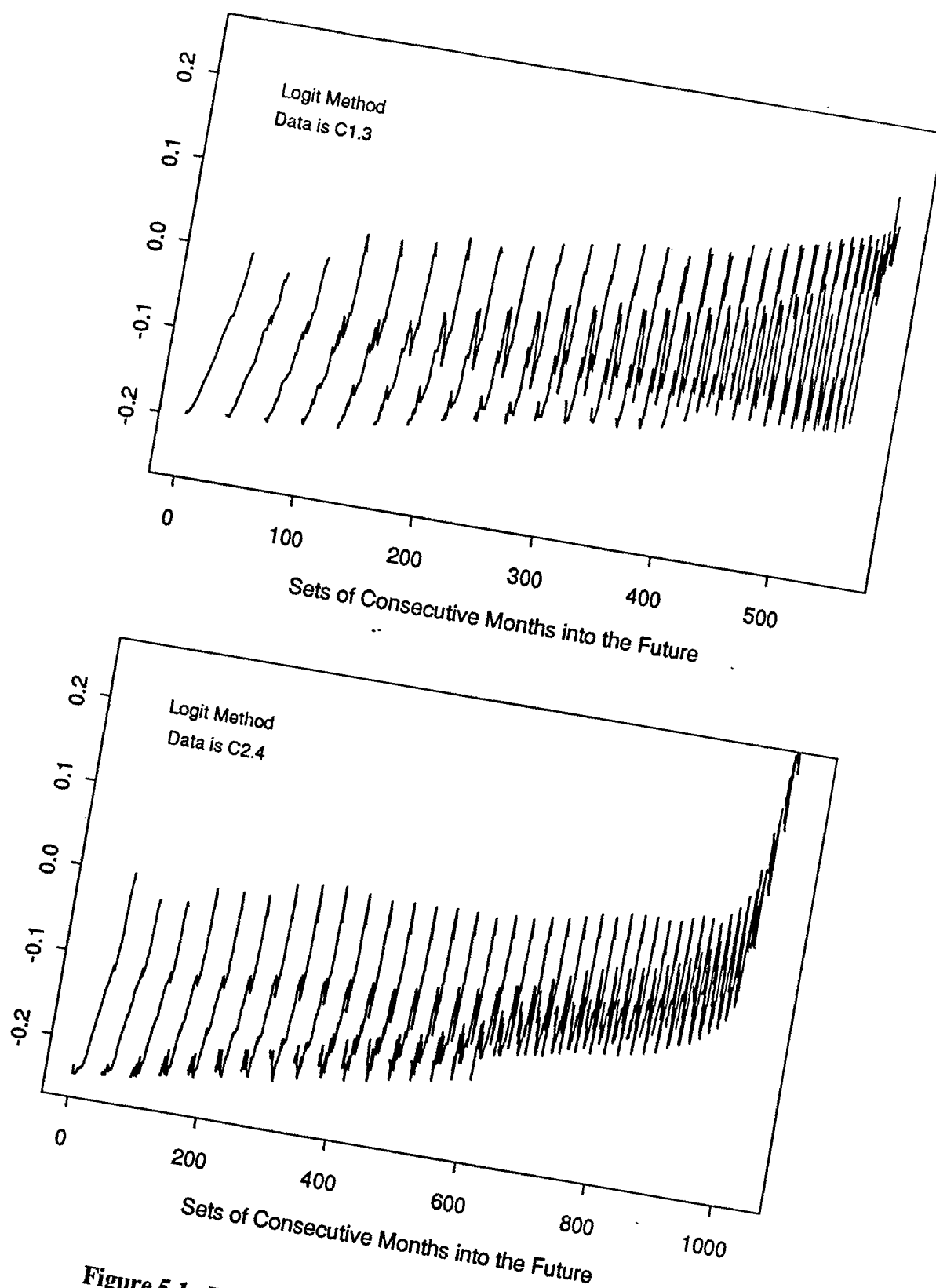


Figure 5.1. Relative Error Graphs with Covariates

We do not have explanations for the behavior, only speculations. The speculations include the following:

- a. It is well known that maximum likelihood estimates are biased for small samples. This effect may be present because the attrition counts are small relative to the inventory counts. Also the latter are quite variable. Treatment for this problem has been developed, see [McCullagh, ch 7], but such has not been studied in the present work.
- b. The modeling has a meta analysis aspect to it that could be mischievous. The parameter estimates are developed from sets of cohorts while the consensus hazard values are developed from the constant s baseline values that cross cut the cohorts. It would be necessary to search the literature further to obtain views on this point.
- c. More careful, and perhaps alternative, model building may improve the situation.

6. SUMMARY

Although the study lacks a completeness in terms of resolving the manner of using covariates, a number of important things have been learned.

- a. There is a need for a clear standard for what is meant by "error of forecast."
- b. The forecasting of attritions is quite sensitive to the partitioning of the estimation data base. Certainly the C -group by contract term is important and there may be other features of the data that are deserving of separation.
- c. The methods studied for obtaining consensus hazard values using averages of fixed mis values crosscutting the cohorts appear to be defensible. The logit, probit, and exponential smoothing methods work about equally well, at least in the no covariate cases using involuntary losses.
- d. The use of generalized linear models with covariates has not found success. This is an enigma. They should be expected to perform at least as well as the methods used in the no covariate case. Perhaps the covariates should be treated as random effects.
- e. The data are sparse in the smaller C -groups and the programs prepared often break down in execution due to lack of sufficient information.

REFERENCES

- ... SPLUS statistical software system, StatSci Div, MathSoft Inc. Seattle.
- Agresti, A. (1990). *Categorical Data Analysis*, Wiley.
- Allison, P.D. (1995). *Survival Analysis Using the SAS System*, SAS Institute, Inc.
- Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall.
- DeWald, E.T. (1996). "A time Series Analysis of U.S. Army Enlisted Force Loss Rates," Master's Thesis, U.S. Naval Postgraduate School, Monterey
- Dillaber, K. (1995). ELIM-COMPLIP, unpublished system summary maintained by ODCSPER's Female Enlisted Loss Inventory Model (FELIM) representative.
- General Research Corporation (1985). ELIM-COMPLIP Technical Supplement," Report 1406-04-85-TR.
- General Research Corporation (1989). ADP Support for Enlisted System: ELIM-COMPLIP System Specification.
- Greenston, P., Mackin, P. and Hogan, P.F. (1996). Econometric Module for ODCSPER Manpower Models, Executive Briefing Slides, Army Research Institute.
- Kalbfleisch, J.D. & Prentice, R.L. (1986). *The Statistical Analysis of Failure Time Data*, Wiley.
- McCullagh, P. (1987). *Tensor Methods in Statistics*, Chapman & Hall.
- McCullagh, P. and Nelder, J.A. (1991). *Generalized Linear Models*, Chapman & Hall.
- Smith, D.A., Eichers, D., Rose, D.E. Jr., and Rostker, D. (1994). "Model Description and Proposed Application for the Enlisted Personnel Inventory, Cost, and Compensation Model," Study Report 94-03, U.S. Army Research Institute.

APPENDIX A **C-Group Structure and Distribution of Accessions**

The C-group definitions are recorded in the table.

Currently Defined Characteristic Groups (GRC; 1995)

CG	Gender	Education	AFQT Category	Term
1	M	HSDG	I-III A	3, 4
2	M	HSDG	IIIB	3, 4
3	M	HSDG	IV-V	3, 4
4	M	NHSDG	I-III A	3, 4
5	M	NHSDG	IIIB-V	3, 4
6	F	HSDG	I-III A	3, 4
7	F	HSDG	IIIB-V	3, 4
8	F	NHSDG	I-V	3, 4
9	M	HSDG & NHSDG	I-V	2, 5, 6
10	F	HSDG & NHSDG	I-V	2, 5, 6

TABLE KEY:

CG: Characteristic Group

Gender: Male (M), Female (F)

Education: High School Degree (HSD), No High School Degree (NHSD)

AFQT Category: I-III A 99-50 percentile

IIIB 49-39 percentile

IV 30-21 percentile

V 0-20 percentile

Term: Length of enlistment contract, in years

For the present research, C-groups 1, 2, and 4 have been downloaded for analysis on a PC. There are two files for each: three year term and four year term. In forming these the V affix (for variable enlistment due to training time) has been ignored. It is useful to record some basic statistics for these data. Below are the annual accession counts for the years 1983 through 1989, and the distributions of these accessions over the twelve months. These distributions are also presented in graphical form. The annual cycles appear to change with the C-group.

Annual Accessions for our Six Data Sets

Year	C1.3	C1.4	C2.3	C2.4	C4.3	C4.4
1983	18968	24362	18335	7424	7758	1309
1984	14956	24555	20250	5017	7111	1703
1985	12993	22954	20073	4879	6323	1530
1986	19608	24532	25266	5879	6062	1368
1987	18144	20683	14772	10237	4156	4124
1988	8165	25803	6248	17037	362	1267
1989	6113	20833	3454	16647	1083	2428

Monthly Distribution of Accessions C1.3

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1983	0.0575	0.0786	0.0984	0.0547	0.0768	0.0505	0.0974	0.1271	0.1104	0.0826	0.1140	0.0519
1984	0.0980	0.0959	0.0869	0.0355	0.0526	0.0819	0.0836	0.1518	0.1022	0.0874	0.0855	0.0388
1985	0.1044	0.0587	0.0575	0.0312	0.0429	0.0809	0.1458	0.1198	0.0900	0.1308	0.1041	0.0339
1986	0.0866	0.0750	0.0566	0.0552	0.0647	0.0615	0.1336	0.1037	0.0866	0.1380	0.0975	0.0410
1987	0.1079	0.0907	0.0702	0.0786	0.0837	0.0722	0.1269	0.1027	0.1133	0.0688	0.0541	0.0308
1988	0.0957	0.0758	0.0866	0.0298	0.0770	0.0764	0.1056	0.1448	0.1334	0.0708	0.0878	0.0164
1989	0.0949	0.0566	0.0509	0.0337	0.0469	0.0584	0.1265	0.1979	0.1209	0.1024	0.1081	0.0028

Monthly Distribution of Accessions C2.3

	[.1]	[.2]	[.3]	[.4]	[.5]	[.6]	[.7]	[.8]	[.9]	[.10]	[.11]	[.12]
[1.]	0.0410	0.0845	0.0774	0.0604	0.0644	0.0509	0.1057	0.1316	0.1347	0.1226	0.0776	0.0492
[2.]	0.0839	0.1019	0.0779	0.0487	0.0628	0.0765	0.1080	0.1442	0.0931	0.0785	0.0914	0.0331
[3.]	0.1056	0.0760	0.0667	0.0406	0.0565	0.0764	0.0963	0.1210	0.0762	0.0965	0.1342	0.0540
[4.]	0.1098	0.0936	0.0561	0.0598	0.0683	0.0514	0.1193	0.1450	0.1467	0.0454	0.0656	0.0391
[5.]	0.0707	0.1032	0.0635	0.0829	0.0790	0.0847	0.1809	0.1408	0.1219	0.0231	0.0249	0.0244
[6.]	0.1220	0.0645	0.0523	0.0407	0.0672	0.1247	0.1373	0.1448	0.1229	0.0343	0.0683	0.0210
[7.]	0.0877	0.0811	0.0420	0.0321	0.0371	0.0654	0.1228	0.1743	0.0831	0.1500	0.1222	0.0023

Monthly Distribution of Accessions C4.3

	[.1]	[.2]	[.3]	[.4]	[.5]	[.6]	[.7]	[.8]	[.9]	[.10]	[.11]	[.12]
[1.]	0.0482	0.0647	0.0838	0.0956	0.0396	0.0534	0.1589	0.1494	0.0837	0.0327	0.1471	0.0429
[2.]	0.1357	0.0918	0.1069	0.0811	0.1187	0.0523	0.0530	0.0882	0.0391	0.0956	0.0886	0.0489
[3.]	0.1485	0.1004	0.1031	0.0716	0.0754	0.0440	0.0655	0.0748	0.0399	0.1085	0.1069	0.0614
[4.]	0.1323	0.1165	0.1024	0.1028	0.0952	0.0092	0.0742	0.0214	0.0493	0.1775	0.0892	0.0299
[5.]	0.1653	0.0974	0.0671	0.0854	0.1080	0.0272	0.0387	0.0255	0.0310	0.2329	0.0996	0.0217
[6.]	0.1492	0.1298	0.0552	0.0552	0.0994	0.1547	0.0967	0.1133	0.0387	0.0387	0.0525	0.0166
[7.]	0.0896	0.1108	0.0674	0.0831	0.1256	0.0757	0.1560	0.0480	0.0148	0.1597	0.0683	0.0009

The above are for the three year contract term. It is interesting that the four year term distributions match their three year term counterparts rather well. Graphs of the above distributions provide a sense of the cycles.

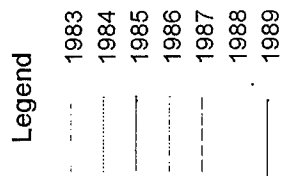
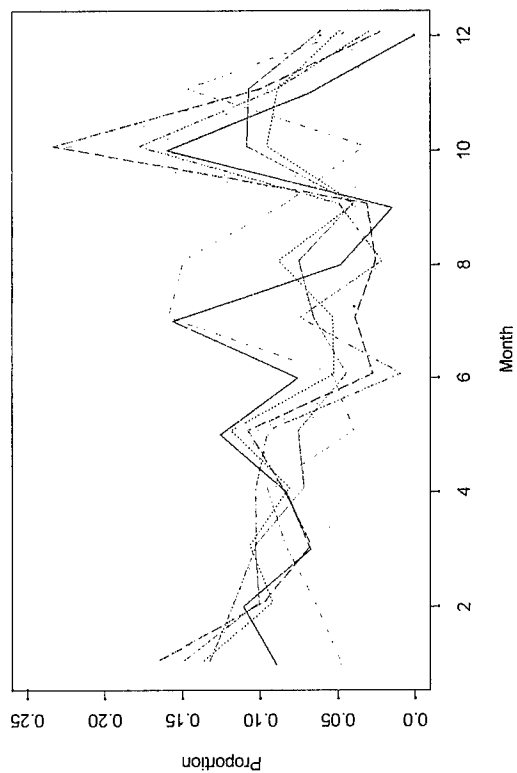
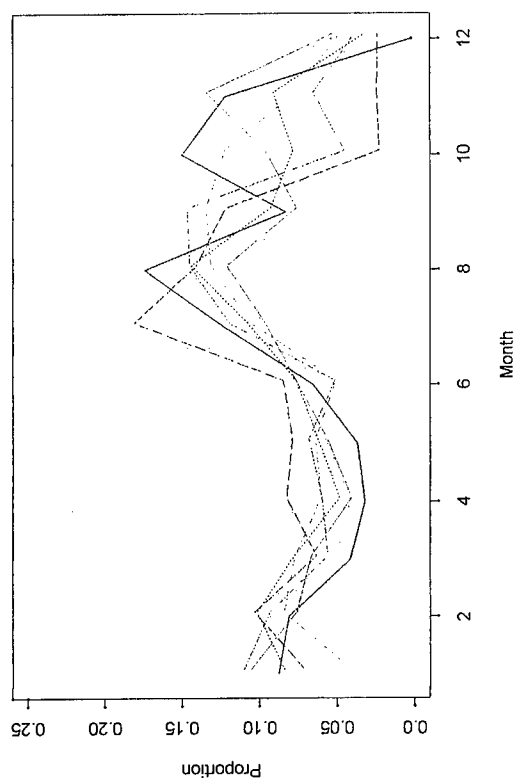


Figure A.1. Annual Distribution of Accessions

APPENDIX B

Data cartridges extracted from the Small Tracking File were supplied by General Research Corporation. The records begin in January of 1983 and extend through December of 1994. These files contain records for individual soldiers and the fields are coded according to the following format.

Cohort	YYYYMM	
SSN	9.	
AFQT	2.	
Race	1.	
Gender	\$1.	M/F
Term	1.	
Cived	\$1.	
Age_at_entry	3.1	
Base_Pay_Entry_Date	YYYYMM	
ETS_Date	YYYYMM	
Component	\$1.	
Original_BASD	YYYYMM	
Current_BASD	YYYYMM	
Filler	\$2.	
Training_Time	2.	
VEL_Flag	\$1.	
Flags	4.	
#Events	2.	
(Month_of_Service	3.	Event \$3. (occurs #Events times)

This file, which comes from the ELIM historical database, tracks individual NPS gains for up to six years through multiple events, such as ETS loss followed by G90 gain and extension to a second term after reenlistment. It therefore tracks soldiers past the point where they become non-C-group.

Table B.1. Required Loss, Reenlistment, Extension (LRE) Rates

CATEGORY OF LRE BEHAVIOR	DATA CODES ON THE SMALL TRACKING FILE (STF)
Drop From Rolls	DFR
Total Adverse	EDP, MCD, TDP, UFT
Entry Level Adverse	TDP
Unsatisfactory Performance	EDP
Other Adverse	MCD, UFT
Physical Disqualification/Disability	PHY
Expiration of Term of Service	ETS
Immediate Reenlistment	IMR
Extension	EXT
Other Loss	BLK, EMP, ERL, HRD, LLL, MPP, OSR, OTH, SCH
Retirement	RET

Generally the file called “dat” in what follows may be thought of as being in the (r, s) plane (see section on indexing conventions) and contain the “at risk” or beginning period inventory counts. The rows are by cohorts and the columns are months in service. There are some enhancements however. Each cohort is repeated for every combination of covariate factor levels. In the present study the code accepts six age groups (less than or equal to 18, over 18 but less than or equal to 20, over twenty but less than or equal to 22, over 22 but less than or equal to 24, over 24 but less than or equal to 28, and over 28) and three racial groups (white, black, other*). Thus the total number of rows should be no more than 18 times the number of cohorts (missing cases excepted) and at least m (cohort length in months) columns. Thus the file dat may be excessive but need not be trimmed to fit the immediate application.

The following few lines illustrate the structure. The atrisk (beginning inventory) counts by cohort and factor level groups for the first C-group, three year enlistment begin as follows.

Table B.2. Structure of Input Data

Cohrt	Age	race	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10...
1. 198301	1	1	11	11	10	9	9	9	8	8	8	8...
2. 198301	1	2	1	1	1	1	1	1	1	1	1	1...
3. 198301	2	1	417	416	405	388	384	382	377	376	376	376...
4. 198301	2	2	41	41	38	36	36	36	36	35	35	35...
5. 198301	2	7	20	20	18	17	16	16	16	.15	15	15...

* If a record has race marked as unknown, it was placed in the ‘other’ group.

APPENDIX C

Exponential Smoothing Technique

The following is our understanding of the method in current use. No use is made of covariates. Use is made of the cells in the data template. The Kaplan-Meier estimates are placed in the cells, i.e.,

$$\tilde{h}(r, t) = \frac{w(r, t)}{\text{Inv}(r, t-1)} \quad \text{for all } (r, t) \quad (\text{C.1})$$

in the data template and $w(r, t)$ is the number of attritions in cell (r, t) , and $\text{Inv}(r, t-1)$ is the number of personnel in cohort r that survive time period $t-1$. These quantities are subject to the exponential smoothing algorithm.

Let us first describe the algorithm and then apply it to the above. Consider a time series

$$x_1, x_2, \dots, x_t, \dots$$

Those values for $t = 1, \dots, n$ have been observed and it is desired to forecast

$$\hat{x}_{n+1}.$$

Given a constant a ($0 < a < 1$), the one step ahead forecast is given by

$$\begin{aligned} \hat{x}_{n+1} &= a x_n + (1-a) \hat{x}_n \\ &= a \sum_{j=0}^{\infty} (1-a)^j x_{n-j}. \end{aligned}$$

Of course one cannot literally sum to infinity so one must choose a finite starting point; $j = n-1$ is the largest that can be used, but it will hardly be noticed if a smaller value is substituted. Indeed, for values of a near unity, this forecast is virtually that of persistence, i.e., $\hat{x}_{n+1} = x_n$. If the above series is summed to $j = k$ then the weights, $a(1-a)^j$, should be normalized by their total, i.e., $[1 - (1-a)^{k+1}] / a$.

For our application this rule is applied to the counterdiagonals of the data template. That is, the values $\{x_t\}$ are replaced by

$$x_t = \tilde{h}(t+1-s, t) \text{ for each } s = 1, \dots, m. \quad (\text{C.2})$$

(Note the initial value of t is not unity but rather a function of s .)

One obtains the forecast $\hat{x}_{n+1}$ for each s . I.e.,

$$\hat{h}(n+2-s, n+1) \quad \text{for each } s = 1, \dots, m,$$

and these serve as the hazard rates for all (s, t) cells in the forecast region using the $\hat{\text{Inv}}(r, t)$ update system described in Appendix F.

APPENDIX D

Counting Formula

It is useful to record formulae for counting subsets of cells in the data template. The following relationships have been developed:

Estimation data cells exist for $r = 1, 2, \dots, n$ and $t = r, r + 1, \dots, \min(n, m + r - 1)$.

The number of cells removed in the upper right corner of the estimation set (whenever $m < n$) is $c_0 = \binom{n+1-m}{2}$.

The total number of data cells to the left of the 'wall' is $N = \binom{n+1}{2} - c_0$.

The month in service index s is a function of (r, t) ; specifically $s = t + 1 - r$.

For fixed r , the range of s is $s = 1, 2, \dots, \min(m, f(r, t))$ where $f(r, t) = \min(t, n + 1 - r)$.

Any cell can be located by specifying any two of the indices selected from (r, t, s) .

Notice the potential size of the setting. If $n = 72$ and $m = 45$ then

$$c_0 = \binom{28}{2} = 378$$

$$N = \binom{73}{2} - c_0 = 2628 - 378 = 2250.$$

Turning to the 'validation region', the number of diamond cells in each column is m , and the total number of diamond cells is m^2 . The number of cells in the critical triangle is $m(m + 1)/2$.

There may be need to index the estimation cells with a single subscript $j = 1, \dots, N$. This will be done by catenating each row to the right of its preceding row. Thus $j = j(r, t)$ is specified by

$$j = (r-1)m + 1 + t - r \quad \text{for } r = 1, \dots, n+1-m$$

$$j = n(r-1) - \binom{r-1}{2} - c_0 + 1 + t - r \quad \text{for } r = n+2-m, \dots, n.$$

A bit of searching must be done to invert this process, i.e., recover the pair (r, t) from j . The smaller values of j are associated with the upper parallelopiped in the template. Here the largest value of t is $m + r - 1$ as a function of the row. This allows one to express a relationship between s and r in this region. Clearly

$$j \leq mr \leq m(n + 1 - m).$$

If j satisfies the far right inequality then we seek r such that $mr \leq j < m(r + 1)$. Having identified r we compute $t = r + (j - mr)$.

It is more complicated when $j > m(n + 1 - m)$. We seek the largest r such that

$$r(n-1)+1-\binom{r-1}{2} \leq j + c_0.$$

Then $t = r$ plus the excess of the right hand side over the left hand side of this inequality.

APPENDIX E

Program Suite Contents

Several SPLUS program suites were developed in the course of the research. All presume the same input data files as described in Appendix B; that is, the 'atrisk' values preceded by cohort and covariate information. What follows are the contents and function names. SPLUS has a large number of generic functions which are used, but no listing of their names is made. The suites of programs are available on disk from the author. The disk contains the functions in an ASCII format which contains some overhead code that allows them to be loaded into any SPLUS installation using the generic function "data.restore".

1. Without covariates.

The tables of Section 4 were computed using the function 'drive.2' and functions called by it. The invoking statement is `drive.2(dat, m, act, a)` where `dat` is the input inventory count file; m is the cohort length parameter; `act` is an m -by m lower triangular matrix of actual 'at risk' counts in the critical triangle; and `a` is the constant for the exponential smoothing algorithm. The array 'act' is obtained by executing the function `actuals(tcode, dat, m)` where `tcode` is the date of the first column beyond the 'wall' in the critical triangle in 198xxx format. The support functions of `drive.2` are named

`kaplan; sam; probit; logmlog; expsm; lin.ave; lin.wtd; valid0; critt0; relerr.`

The output of `drive.2` is a list that contains the predictions made by the six methods mentioned in Section 4.

2. With covariates.

The computations summarized in Section 5 were created from executing the function `driver(dat, r, k0, m, type1, type2, tcode)`. The input `dat` is the same as described; r is the number of cohorts in an echelon; k_0 is the number of echelons in the estimation stage; `type1` is a character variable specifying the estimating hazard function data source (choice is "current" or "grandm"); `type2` is a character variable specifying the consensus hazard estimation method (choices are "logit", "compll", "probit", "linear"); and `tcode` is the beginning date of the critical triangle as before in 198xxx format. The six functions in the direct statement of `driver` are

`coef.est; hazarray; critta; valid1.set; convert; forecast.`

Other functions in support are: adjust; block.3; block; prep.dat; baseline.ind;
 modadj; base.haz; hazwall; yrtomo; motoyr.

The output is a lower triangular (order m) matrix of predicted at risk counts for the critical triangle.

A plotting function, relplot, produces the graphs when passed the output of relerr. The C-group-1 data set for 3 year term is included on the disk.

APPENDIX F

General Structure of Estimation and Forecasting

This appendix contains the abstract structure of attrition estimation using generalized linear models. Further details can be found in [McCullagh & Nelder]. What appears here is the method of blending those models with the special needs of the present cohort based data structure.

Estimation of Parameters

Let $q(r, t)$ represent the baseline hazard function estimates produced by the Survival Analysis. The indices (r, t) will have to be organized into a single subscript $j = 1, \dots, N$ where N is the number of cells in the data template. We indulge in the notation conventions exemplified by

$$q_j = q(r, t), \text{ etc.}$$

and use will be made of either when convenient.

The model used for forecasting utilizes a Generalized Linear Model (GLIM). Specifically:

- i) There is a set $\mu(s)$ of consensus hazard rates for $s = 1, \dots, m$.
- ii) The hazard rate for a specific (r, t) cell is found from $\mu(s)$ (for $s = t + 1 - r$) modified by appropriately selected covariates.

A GLIM with transform will be executed in order to estimate the $\{\mu(s)\}$.

Apply the selected transform ϕ to the $\{q_i\}$ to produce y , an N component response vector whose j^{th} component is

$$y_j = \phi(q_i) \quad j = 1, \dots, N. \quad (\text{F.1})$$

Then the basic linear model is

$$y = X\beta + e \quad (\text{F.2})$$

where the matrix X is an N by p matrix of explanatory variables and the error term has covariance structure proportional to V , an N by N matrix. The normal equations are

$$(X'V^{-1}X)\hat{\beta} = X'V^{-1}y. \quad (\text{F.3})$$

Generally, the matrix V must be updated with each solution. I.e., apply iteratively reweighted least squares.

The raw fitted values of the hazard rates are found by inverting the transform ϕ

$$\hat{q}(r, t) = \phi^{-1}(\hat{y}_j). \quad (\text{F.4})$$

These are the fitted baseline hazard rates. For purposes of forecasting they must be modified by the covariates used in the proportion hazards development:

$$\hat{h}(r, t) = e^{z'\beta} \hat{q}(r, t) \quad (\text{F.5})$$

or the proportional odds development (eq (10)) as appropriate, where $z'\beta$ is the appropriate linear form for cell (r, t) .

Validation

1. Data: The data reserved for validation, referring to the template, are those cells indexed by

$$\begin{aligned} r &= n + 2 - m, \dots, n \\ t &= n + 1, \dots, n + m - 1 \text{ and } t \leq n + r - 1. \end{aligned}$$

For now we limit ourselves to those cohorts in the critical triangle.

We require the inventory counts entering all cells with $t = n + 1$ (that is, the number that survive time n for each $r = n + 1 - m, \dots, n$) broken out by covariates. The covariate system will be denoted by θ ; each member of θ is a covariate combination. We also require the number of losses in these cells also broken out by covariates. I.e., the set

$$w_{\theta}(r, t) = \text{number of attritions in cell } (r, t) \quad (\text{F.6})$$

for each component of θ . Similarly, the number that survive the n^{th} time period (inventory entering $t = n + 1$) will be denoted

$$\text{Inv}_{\theta}(r, t) \text{ evaluated at } t = n. \quad (\text{F.7})$$

These inventories are needed to convert attrition rate estimates into attrition forecasts, $\hat{w}_{\theta}(r, t)$.

The error of forecast is the value of the actual removed from the forecast,

$$\text{forecast error} = w_{\theta}(r, t) - \hat{w}_{\theta}(r, t) \text{ summed over the members of } \theta.$$

Consider any estimator of the hazard function

$$\hat{h}_{\theta}(s, t) \text{ with } s = t + 1 - r.$$

This is the estimated attrition probability s periods into the cohort and at real time t . One estimates the number of losses using

$$\hat{w}_{\theta}(r, t) = \hat{\text{Inv}}_{\theta}(r, t-1) \hat{h}_{\theta}(s, t) \quad (\text{F.8})$$

where

$$\hat{\text{Inv}}_{\theta}(r, t) = \hat{\text{Inv}}_{\theta}(r, t-1) - \hat{w}_{\theta}(r, t) \quad (\text{F.9})$$

and the estimated inventories are completed recursively, starting with $t = n$ and the initializing statement

$$\hat{\text{Inv}}_{\theta}(r, n) = \text{Inv}_{\theta}(r, n).$$

That is, begin with the known inventory that enters the $n + 1^{\text{st}}$ period.

INITIAL DISTRIBUTION LIST

1. Research Office (Code 09) 1
 Naval Postgraduate School
 Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013) 2
 Naval Postgraduate School
 Monterey, CA 93943-5002

3. Defense Technical Information Center 2
 8725 John J. Kingman Rd., STE 0944
 Ft. Belvoir, VA 22060-6218

4. Therese Bilodeau 1
 Dept of Operations Research
 Naval Postgraduate School
 Monterey, CA 93943-5000

5. Prof. Robert R. Read (Code OR/Re) 1
 Naval Postgraduate School
 Monterey, CA 93943-5000

6. Prof. Siriphong Lawphongpanich (Code OR/Lw) 1
 Naval Postgraduate School
 Monterey, CA 93943-5000

7. Captain J. Crino 1
 Deputy Chief of Staff, Personnel
 U.S. Army
 Rm. 2C744
 300 Army, Pentagon
 Washington, DC 20310-0300