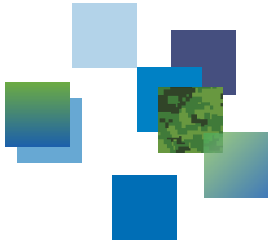
Defence Research and Development Canada
Recherche et développement pour la défense Canada

DRDC | RDDC



A survival analysis of ADM (Materiel) workforce attrition

Examining attrition phenomenon through non-parametric and semi-parametric survival methods

Christopher E. Penney
DRDC – Centre for Operational Research and Analysis

Defence Research and Development Canada

Scientific Report
DRDC-RDDC-2016-R162
October 2016

Canada

A survival analysis of ADM (Materiel) workforce attrition

Examining attrition phenomenon through non-parametric and semi-parametric survival methods

Christopher E. Penney
DRDC – Centre for Operational Research and Analysis

Defence Research and Development Canada
Scientific Report
DRDC-RDDC-2016-R162
October 2016

Abstract

This paper explores the problem of workforce attrition and the role of economic factors in attrition behaviour. There are two main elements to this study: first, I examine the overall attrition phenomenon using non-parametric and semi-parametric models; second, I consider the role of economic factors by limiting the sample to individuals for whom complete histories are available and then employing a semi-parametric model that allows for time-varying elements.

The findings for the primary analysis indicate a strong model fit can be obtained through the inclusion of a few explanatory variables: these are 'starting age', 'starting pensionable years of service', and 'gender'. The findings for the secondary analysis provide some evidence of a role in attrition behaviour for economic factors such as 'salary' and 'classification level', but the effects are offsetting and difficult to distinguish statistically.

I recommend that the client provide data with a significantly longer time horizon so that a more precise determination of the role of economic and financial factors can be obtained.

Résumé

Dans le présent document, j'explore le problème de l'attrition des effectifs ainsi que le rôle des facteurs économiques dans la tendance en matière d'attrition. L'étude comporte deux grands volets : dans un premier temps, j'examine le phénomène de l'attrition dans son ensemble au moyen de modèles non paramétriques et semi paramétriques ; dans un deuxième temps, j'étudie le rôle des facteurs économiques en utilisant un échantillon composé uniquement de personnes dont l'historique complet est accessible, puis en employant un modèle semi paramétrique permettant la prise en compte d'éléments variant dans le temps. Les résultats de l'analyse initiale indiquent qu'il est possible d'obtenir un modèle bien ajusté grâce à l'inclusion de quelques variables explicatives, à savoir : l'âge au début, le nombre d'années de service ouvrant droit à pension au début et le sexe. Les résultats de la seconde analyse révèlent que certains facteurs économiques comme le salaire et le niveau de classification jouent un rôle dans la tendance en matière d'attrition, mais que leurs effets sont conflictuels et difficiles à cerner statistiquement parlant. Je recommande au client de fournir des données avec un horizon temporel beaucoup plus long de façon à pouvoir faire une évaluation plus précise du rôle des facteurs financiers et économiques.

Significance for defence and security

The client, Directorate Materiel Group Management Coordination (DMGMC), has requested that the Defence Economics Team (DET) undertake an empirical investigation into the problem of workforce attrition for members of the Purchasing and Supply (PG) group within the Department of National Defence, with attention towards the role of economic factors. As hiring suitable candidates for the public service is both time consuming and costly, accurate models of attrition should be employed in order to develop predictions of workforce levels and inform future recruiting decisions.

Using a survival analysis approach to study the attrition phenomenon, this study finds:

- A simple semi-parametric model provides a strong fit of the data. This methodology can be used to develop predictions of workforce levels.
- A more sophisticated semi-parametric model with time-varying covariates indicates that individual economic factors may have an impact on the likelihood of attrition due to outside employment, but these factors have competing effects that render precise statistical inference problematic.
- I recommend that the client provide data with a longer time horizon, perhaps 10 additional years, to more precisely examine the role of economic factors in workforce attrition.

Importance pour la défense et la sécurité

Le client, le Directeur – Coordonation de la gestion du Groupe des matériels (DCGGM), a demandé à l'Équipe de l'économie de la Défense (EED) de réaliser une étude empirique sur le problème de l'attrition des effectifs au sein du groupe Achats et Approvisionnements (PG) du ministère de la Défense nationale en portant une attention particulière au rôle des facteurs économiques. Comme le recrutement de candidats qualifiés dans la fonction publique est un exercice à la fois long et coûteux, il est recommandé d'utiliser des modèles d'attrition précis pour la prédiction des niveaux d'effectifs et faciliter la prise des décisions futures en matière de recrutement. L'application de la méthode de l'analyse de survie à l'étude du phénomène de l'attrition a permis de faire les constatations suivantes :

- Un modèle semi-paramétrique simple donne un bon ajustement des données. Cette méthodologie peut être utilisée pour la prédiction des niveaux d'effectifs.
- Un modèle semi-paramétrique plus sophistiqué avec des variables évoluant dans le temps permet d'établir les facteurs économiques individuels qui peuvent influer sur la probabilité d'attrition due à l'embauche externe, mais ces facteurs peuvent avoir des effets conflictuels compliquant l'établissement d'une inférence statistique précise.
- Je recommande au client de fournir des données avec un horizon temporel plus long, disons dix années de plus, de manière à permettre un examen plus précis du rôle des facteurs économiques dans l'attrition des effectifs.

This page intentionally left blank.

Table of contents

List of figures

List of tables

1 Introduction

The client, Directorate Materiel Group Management Coordination (DMGMC), has requested that the Defence Economics Team (DET) undertake an empirical investigation into the problem of workforce attrition for members of the Purchasing and Supply (PG) group within the Department of National Defence. The client also seeks an understanding on the role of economic factors in attrition behaviour.

Following the example of an earlier analysis conducted by the DET ([1]), I explicitly consider the problem as one within the purview of time-to-event phenomena. In this context, the probability of an individual leaving the workforce is defined as a function of time. This assertion is easily justified: individuals begin their PG careers, remain within the group for some time, and then exit. Certain factors[1] may have an influence on the instantaneous departure probability. Survival analysis methods allow for the study of this class of problem in a variety of rigorous manners, including the use of non-parametric and semi-parametric models. I adopt both of these approaches within this research.

Previous analyses[2] on this phenomenon have adopted alternate approaches, such as using standard panel data regression methods to model attrition behaviour as a "stay or leave" decision that is affected by a set of factors. Survival models, conversely, exploit the inherent censoring information available within time-to-event data, thus avoiding inconsistencies and bias that may result in cases where this information is ignored.

To ensure that this study is both scientifically defensible and capable of providing meaningful insights, I approach the problem from a parsimonious modelling philosophy; that is, by making *as few assumptions as possible*. To this end, I avoid the arbitrary inclusion of covariates by using stepwise statistical criteria in variable selection. Further, the survival methods used in this research impose minimal structure on the underlying data; in particular, I employ the Kaplan-Meier and Cox Proportional Hazards models.

I also conduct a separate analysis to examine the role of economic and financial factors. For this, I draw upon a smaller portion of the population, specifically one that includes only individuals for whom the dataset time frame contains their initial entry into the PG group. This subset therefore allows for the examination of time-varying economic and financial factors. The type of attrition examined in this model is different than in the main sample: using the 'leave reasons' available within the dataset, I impose a narrower definition of

[1] In the context of workforce attrition, factors such as *age*, *gender*, and *pensionable years of service*, to name only a few, may influence the attrition likelihood.

[2] See, for instance, [2] and [3].

attrition that focuses only on exits for outside employment. The model I adopt for this portion of the analysis is the Extended Cox Proportional Hazards model.

The remainder of this paper is structured as follows: Section 2 describes both the primary population and the "full histories" sample used for the secondary analysis. Section 3 discusses the study's methodology. Section 4 presents the model results and Section 5 concludes.

2 Data

Personnel and mobility data for members of the PG group was provided to the DET by the Directorate Strategic Planning and Accountability (DSPA) through the client's liaison, the Directorate General Materiel Group Operational Research (DMGOR). To ensure the confidentiality of the personnel, I combined the data sources and applied a random identification number to each individual. The random identification number remains with each member throughout their tenures in the dataset, allowing them to be tracked over time.

I employ two samples in this analysis. The first sample corresponds to the full dataset provided by the DMGOR. The second sample includes only individuals who joined the workforce as PGs after 2002 and thus have observations for every year they are present within the group.

The full dataset has records for the years 2003 to 2014. In total, there are 1,385 individual members of the PG group who are active for at least one period (i.e., present in the dataset for at least one year). There are a total of 8,271 person-years to draw upon within this dataset.

Variables provided by the client include 'age', 'years of service' within the PG group, 'total pensionable years of service' (PYOS), 'gender', 'first language', 'marital status', 'employment status', 'classification level' within the PG group, and 'salary'. As both 'level' and 'salary' require complete histories to be modeled correctly, their use is limited to the secondary analysis.

Other variables considered for this analysis but were not available in the data provided by the client include 'number of dependents' and 'health status'.

2.1 Data Manipulations

The original dataset provided by DMGOR requires cleaning to be suitable for use in regression analysis. To correct errors in the data, I assume that the initially reported values for each individual are accurate, and produce imputations for any subsequent missing or erroneous data. For example, since the values for the 'age' of individuals often fail to increment along with calendar years, or increment twice in the space of one year, I use the first reported value of 'age' and increment by one unit for each year. I also assume variables that do not vary over time, such as 'gender' and 'first language', remain constant, and carry out imputations accordingly.

For the post-2002 sample, I transform salaries in each time period to reflect *2010 dollars*, which allows for comparison over time. In the event of a missing or erroneous value for the 'levels' variable, I impute the last reported value.

2.2 Full Sample

Table 1 presents the descriptive statistics for the full sample. Of the 1,385 individuals in the population, 546 leave the service before the end of the observed time (2014) for an attrition rate of 39.4% over the 12-year period.

The PG group contains many individuals who are former members of the military. These individuals tend to start with a much higher level of 'pensionable years of service' and 'age' than others. Thus, the distributions for these two variables are bimodal. This can be more clearly seen in Figure 1. Splitting the sample at the age of 40, we can see that the older group has a substantial proportion of entrants in the vicinity of the 20 pensionable years of service mark, while the younger group is congregated very closely at the 0 pensionable years of service mark.

2.3 Post-2002 Sample

The secondary sample employed in this research consists *only* of those individuals for whom *complete histories*[3] are available. Since the data begins in 2003, this sample includes only those who begin their careers as PGs (i.e., begin with zero years of service in the PG group) from this point onward.

Table 2 lists the summary statistics for the the post-2002 sample. This subsample differs from the full sample in a few areas. Individuals included in the subsample tend to have spent fewer years of service as a PG and are less likely to have retired or otherwise left the service. The complete history nature of the subsample data allows for the inclusion of other variables of interest, including 'salary' and 'level'. The bimodal distribution of 'starting age' and 'starting PYOS' for individuals over the age of 40 is again present in the post-2002 subset of the data, represented in Figure 2.

[3] To clarify, an individual is said to have a 'complete history' if and only if they have been followed within the dataset from the beginning of their career as PGs until present, or until that individual is censored due to an exit.

Table 1: *Descriptive Statistics, PG Group, Full Sample.*

Statistic	Mean	St. Dev.	Min	Max
Years of service	11.91	10.07	0	40
Starting age	34.74	10.54	18	67
Starting PYOS[a]	4.06	8.74	0	55
French[b]	25.0%			
Married	57.3%			
Female	50.0%			
Attrite	39.4%			

(*a*) Pensionable Years of Service; (*b*) French used as primary language
1,385 total observations.

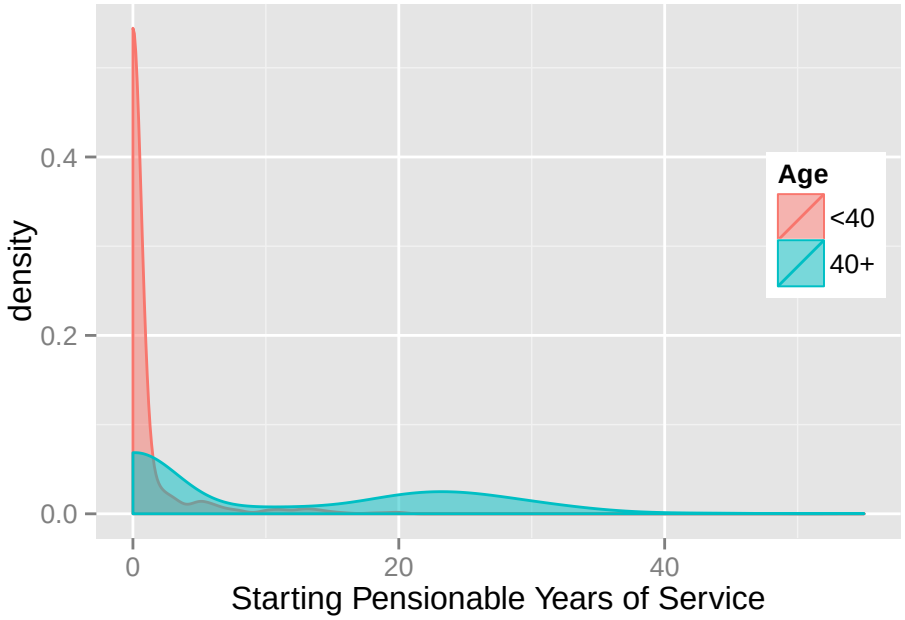


Figure 1: *Starting Age versus Starting PYOS, Full Sample.*

Table 2: *Descriptive Statistics, PG Group, Post-2002 Sample.*

Statistic	Mean	St. Dev.	Min	Max
Years of service	5.95	2.92	0	12
Starting age	37.68	10.36	19	67
Starting PYOS[a]	5.10	9.73	0	55
Starting salary	57,289	12,089	37,500	96,000
Starting level	2[b]	-	1	6
French[c]	24.1%			
Married	56.5%			
Female	47.3%			
Attrite[d]	32.5%			

(*a*) Pensionable Years of Service; (*b*) *Median* start level;
(*c*) French used as first language; (*d*) Includes all leave reasons
932 total observations.

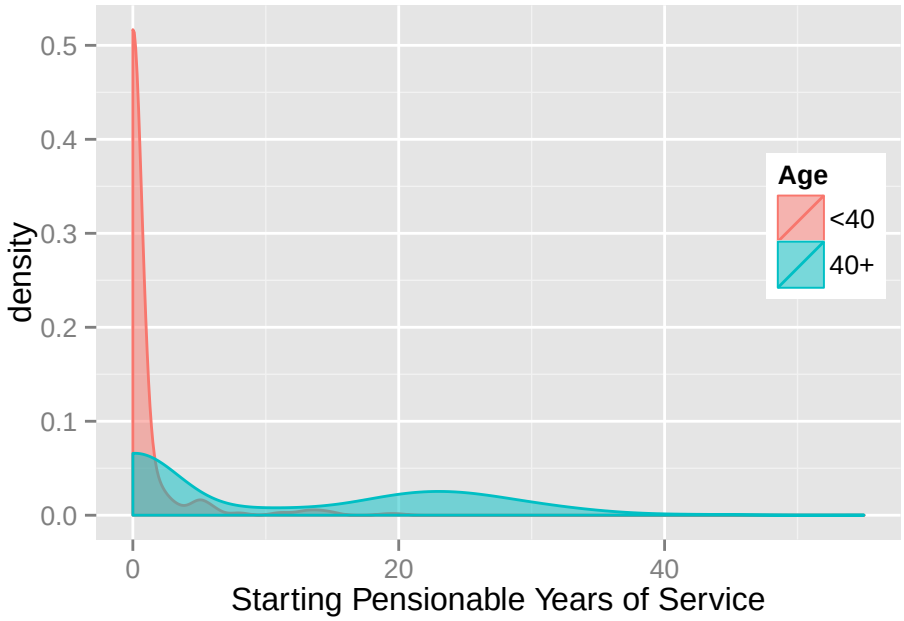


Figure 2: *Starting Age versus Starting PYOS, Secondary Sample (Post-2002).*

3 Methods

"Time-to-event" data refers to data that records the amount of time that passes before an event of interest occurs. In the context of workforce attrition, researchers seek to quantify the risk that an individual may exit the workforce at any given point in time. Survival models exist to handle data of this type.

There are two primary quantities of interest in this survival analysis exercise. The first is the survival function $S(t)$; this describes the probability that a random failure (an exit of the PG workforce) has occurred by time t. At time $t = 0$, $S(t)$ is equal to one, as there is no probability of an exit if no time has elapsed. As t tends to infinity, $S(t)$ approaches zero, as no individual can remain in the workforce indefinitely. The second quantity of interest is the hazard rate $h(t)$, which describes the rate of instantaneous failure (workforce exit). The hazard rate and survival function are linked through the relationship in Equation 1.

$$\hat{S}(t) = \exp\left(- \int_0^t h(s)ds \right). \tag{1}$$

Beginning with the non-parametric Kaplan-Meier estimator, an estimate can be obtained for the attrition hazard based only on the amount of time an individual has existed within the sample.

To examine the effect of covariates, the Cox Proportional Hazards Model provides a framework that allows for variables to shift the underlying attrition hazard; covariates can therefore proportionally increase or decrease the likelihood of instantaneous departure.

One of the primary strengths of survival analysis is its ability to account for statistical problems relating to the censoring of observations. Right censoring is the most common feature of the present data; that is, we expect a given individual will eventually leave the population, but since we have only observed up to a certain time period, we do not always know, for each individual, when that will occur. These individuals are considered 'censored'. The remaining individuals - those who 'fail' within the allotted time as covered within the dataset - are 'uncensored'.

For more information on survival analysis methods, including greater detail on the statistical properties of these models, see [4].

3.1 The Kaplan-Meier Model

The Kaplan-Meier (or Product Limit) Estimator ([5]) represents the standard estimator of the survival function $S(t)$. This non-parametric estimator is a step function with downward kinks at observed failure times.

The Kaplan-Meier method estimates the survival function as follows:

$$\hat{S}(t) = \begin{cases} 1, & \text{if } t < t_1. \\ \Pi_{t_i \leq t}(1 - \frac{d_i}{Y_i}), & \text{if } t_1 \leq t. \end{cases} \tag{2}$$

where d_i represents the number of failures and Y_i represents the individuals still at risk at time t_i.

The Kaplan-Meier model is capable of producing a plot of the survival function generated, including pointwise confidence bands, which provides useful summary information of the attrition behaviour of the population under study ([4]).

3.2 The Cox Proportional Hazards Model

While the Kaplan-Meier estimator is often quite descriptive of the underlying behaviour being studied, it does not take into account the effect that given factors may have on this behaviour. For instance, it is likely that one's age has an effect on how long they will remain in the workforce: if an individual begins their career in the public service at age 20, they would be very likely to remain in the service for a longer amount of time than an individual who joined at the age of 50.

For the examination of these types of factors, I turn to the Cox Proportional Hazards Model (Cox PHM).[4] The Cox PHM allows for the inclusion of covariates that affect the survival function. Specifically, in this framework, covariates are assumed to be multiplicatively related to the hazard rate $h(t)$. The relationship between the covariates and the hazard rate in the Cox PHM is defined as follows:

$$h(t, \mathbf{x}) = h_0(t)\exp(\beta^T \mathbf{x}). \tag{3}$$

where $\mathbf{x}$ represents a vector of covariates included in the model. The function thus causes the underlying hazard to be rescaled according to the estimated values of the coefficients β. In this manner, the effect of a given covariate can be expressed in terms of how it affects the underlying hazard rate $h(t)$.

[4] [6]. Again, for more detail on this class of model, see [4].

3.3 The Extended Cox Proportional Hazards Model

To study the effects of economic and financial factors, it is advantageous to adopt a modelling framework that can account for variables that change frequently with time. The standard Cox PHM requires some generalization in order to properly handle time-varying covariates.

The Extended Cox PHM framework maintains the proportionality assumption introduced previously, and the effects of the covariates, time-varying or otherwise, are interpreted in an identical manner. The model relates the covariates to the hazard rate as follows:

$$h(t,\mathbf{x}) = h_0(t)\exp(\beta_1\mathbf{x}_1 + ... + \beta_k\mathbf{x}_k + \alpha_1\mathbf{v}_1(t) + ... + \alpha_q\mathbf{v}_q(t)). \tag{4}$$

where $\mathbf{x}$ represents covariates $1,...,k$ that do not vary over time, $\mathbf{v}(t)$ represents time-varying covariates $1,...,q$, whose coefficients are represented by α. Note this construction is identical to the standard Cox framework aside from the addition of time-varying covariates $\mathbf{v}(t)$. Thus, both time-invariant and time-varying covariates can be included in the Extended Cox PHM framework.

3.4 Approach to Variable Selection

I adopt a robust variable selection process for the models employed in this study. Naturally, the potential inclusion of any variable comes at a cost of model complexity; to reduce the risk of adding unnecessary covariates to the model, I proceed with a stepwise Akaike Information Criterion (AIC) approach [7]. Stated simply, the AIC provides a measure of the fit of a model given a set of data; adding or removing variables results in changes to the AIC, which informs us as to which model is of higher quality. The stepwise approach begins with a model containing only a baseline hazard rate with no covariates; this is compared to a set of one-variable models, one each for all individual variables included in the dataset. The specification with the best AIC score is then selected. This process continues testing the inclusion of additional variables until the AIC ceases to improve, at which point interaction terms are tested. The final model is reached once the AIC determines that the model does not improve by including any further terms.[5]

[5] A process that begins with a full model including all possible variables and interactions and then successively removing variables is also used within this study, in each case returns the same results as the forward stepwise approach.

4 Analysis and Results

I begin this exercise with summary information on the attrition phenomenon as demonstrated by the Kaplan-Meier estimator.

4.1 The Kaplan-Meier Model

Figure 3 represents the Kaplan-Meier estimate of the survival curve, with 95% confidence bands included. The interpretation of this figure is as follows: the x-axis represents the number of years of service *within the PG group* an individual has attained, while the y-axis indicates the expected survival probability. The survival curve therefore relates the number of attained years of service to a survival probability.

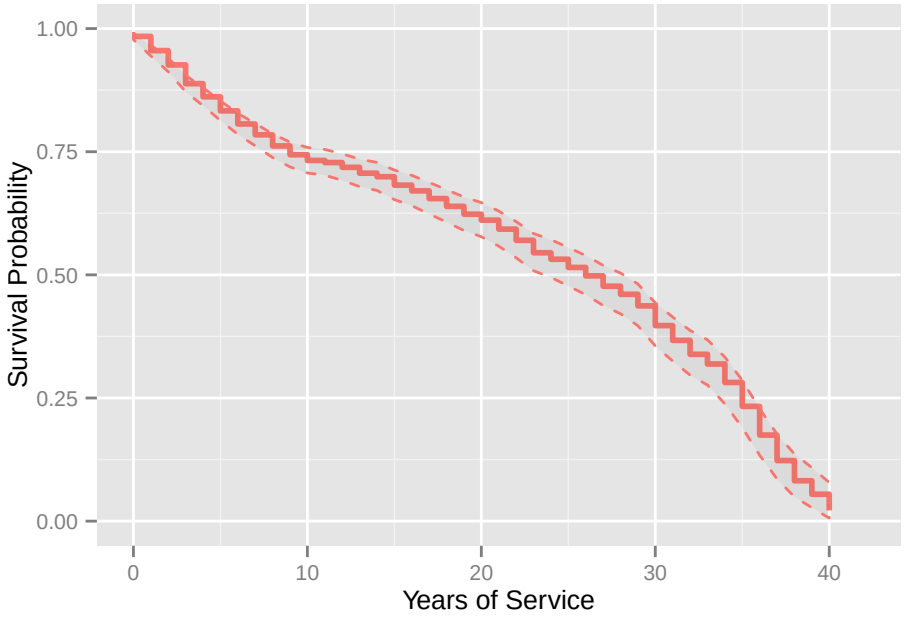


Figure 3: *Kaplan-Meier Curve, Full Population.*

Examining the figure, we can see that the median PG group member will survive until approximately the 27-year mark, with a rapid drop-off following that point. Very few individuals remain in the service beyond a period of 40 years.

The Kaplan-Meier estimate of the survival function is not meant to be able to provide specific insight on given individuals as it is not able to incorporate covariate information.

However, this model makes as few assumptions as possible in constructing the survival curve, and describes the overall phenomenon in a statistically robust manner. To explore the effect of covariates, I proceed with the Cox PHM in the next section.

4.2 The Cox Proportional Hazards Model

The dataset contains several variables that are potentially useful in improving the explanatory power of the model. In particular, these variables are 'starting age', 'starting PYOS', 'gender', and a binary variable indicating whether the individual speaks french as a first language.[6] Some variables, such as 'salary', can only be used in the secondary analysis as they require complete histories to be properly estimated in the survival analysis context.

Proceeding with the stepwise variable selection process described in Section 3.4, the final model is as shown below in Equation 5. The test statistics obtained during the variable selection process are given in Annex A.

$$S(t) = \exp\left(-\int_0^t h_o(s)\exp\left[\beta_1 x_{startage} + \beta_2 x_{startPYOS} + \beta_3 x_{female} + \beta_4 x_{startPYOS*female}\right] ds \right) \tag{5}$$

The specification therefore includes 'starting age', 'starting PYOS', 'female', and an interaction between 'starting PYOS' and 'female'. This last term captures the possibility that females may be more or less likely to attrite as a result of a higher 'starting PYOS' than males.

Table 3: *Cox PHM Results, Full Sample.*

Variable	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Starting age	0.0593	1.0611	0.0056	10.55	0.0000
Starting PYOS	0.0094	1.0094	0.0058	1.620	0.1040
Female	-0.3252	0.7224	0.0992	-3.280	0.0010
St. PYOS · Female	0.0207	1.0210	0.0089	2.340	0.0190

p-values are approximate.

[6] This variable is referred to as 'french' for the remainder of this paper.

Table 3 presents the estimation results. Here, the *coef* and *exp(coef)* columns return the coefficient values obtained by the regression and their respective exponations. The *se(coef)*, *z*, and *Pr(>|z|)* columns list the usual regression diagnostics for each variable.

Both starting age and starting PYOS are highly significant and represent proportional increases in the underlying hazard function. Females, meanwhile, are slightly less likely to attrite on aggregate. The interaction between starting PYOS and female indicates that the positive effect of starting PYOS on the attrition hazard is slightly more pronounced for women than it is for men. Coupled with the baseline decrease in attrition likelihood for females, this coefficient indicates that females who begin their PG careers with less experience are *less* likely to attrite, while those who begin with more experience are *more* likely to attrite. The turning point occurs at a starting PYOS of about 16 years.[7]

4.2.1 Diagnostics

The Cox-Snell residuals for this model are presented in Figure 4. We can assess the overall model fit by observing how closely the residuals, represented in red, follow a 45° line. The results here are suggestive of a good model fit, with relatively few deviations from the line. Despite this, there is some evidence of a violation of the proportional hazards assumption, as shown by the diagnostics in Table 4: the starting age variable returns a very low Grambsch-Therneau p-value ([8]) of 0.0002.[8] One potential explanation for this result is the presence of time-varying relationships between covariates and the survival function: complete histories for all individuals in the dataset would be required to remedy this issue.

Other diagnostics presented in Table 4 are favourable. The Likelihood ratio test, Wald test, and Logrank test all return strong rejections of the null hypothesis that all coefficients are equal to zero.

4.2.2 Comparisons

Figures 5 through 8 demonstrate the effects of each of the model's covariates. The effect of a given covariate is observed by imposing changes upon it while holding the others constant at some specified value.

[7] That is, females have a lower attrition likelihood than males if they begin with 16 or fewer PYOS, and a higher attrition likelihood otherwise.

[8] This p-value indicates a rejection of the null hypothesis that the proportional hazards assumption is not violated.

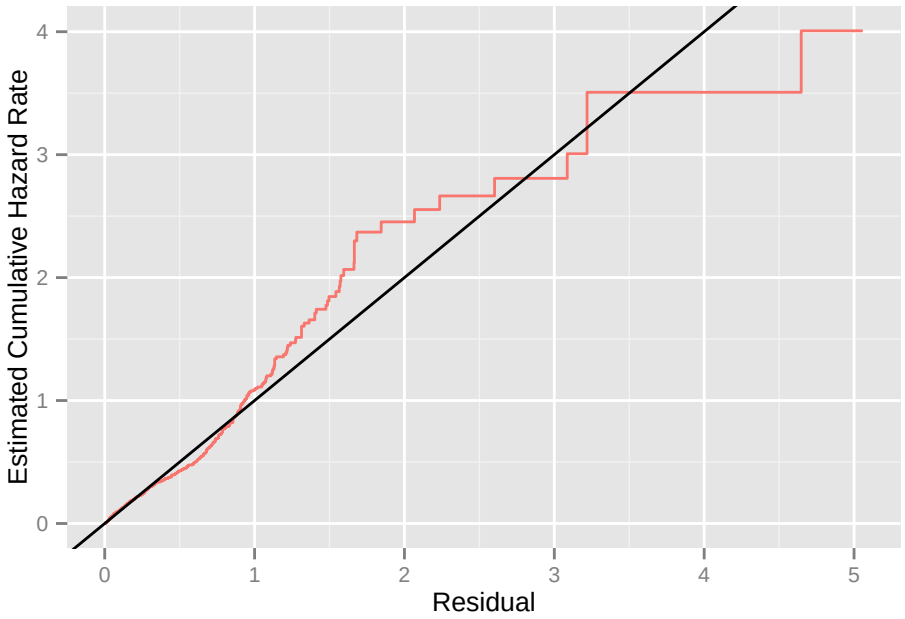


Figure 4: *Cox-Snell Residuals.*

Table 4: *Cox PHM Diagnostics, Full Sample.*

Variable	Pr(>|z|)	Grambsch-Therneau p-value
Starting age	0.0000	0.0002
Starting PYOS	0.1040	0.8916
Female	0.0010	0.7270
St. PYOS* Female	0.0190	0.9452

Concordance = 0.674 (se = 0.015)
R-square = 0.15
Likelihood ratio test = 225.8 on 4 df, p = 0
Wald test = 231.3 on 4 df, p = 0
Score (logrank) test = 258 on 4 df, p = 0

p-values are approximate.

First, Figure 5 compares the survival functions of males and females. The starting age and starting PYOS covariates are held constant at their respective sample means. As indicated in the previous section, females with fewer than 16 starting PYOS are less likely to attrite at any given point than males; since the mean of starting PYOS is 4.06, the overall effect is a decrease in the attrition hazard, represented by an upward shift in the underlying survival curve.

Figure 6 displays the survival curves for individuals with different starting ages. In this scenario, individual is assumed to be male, and the starting PYOS variable is set to zero. This figure clearly shows the downward shifts in the survival function as a result of successive increases in starting age.

The effect of variation in starting PYOS is shown in Figure 7. Since it is impossible for an individual to have more PYOS than, say, their age minus 18, the age covariate has been set to 50 for each of these curves. The female covariate, meanwhile, is held constant at zero. The remaining effect is then purely that of changing starting PYOS; survival curves are drawn for the values of 0, 15 and 30. Again, the graph demonstrates a downward shift for successively higher starting values of PYOS.

Finally, Figure 8 presents two extreme cases. In the first case, an individual begins working as a PG at age 20, with zero prior PYOS; in the second, the individual begins at 50 years of age with 30 PYOS. Both curves have the binary female covariate set to zero. We can see that the latter individual is unlikely to remain in service for even 10 years, while the younger individual is expected to attain roughly 30 years of service.

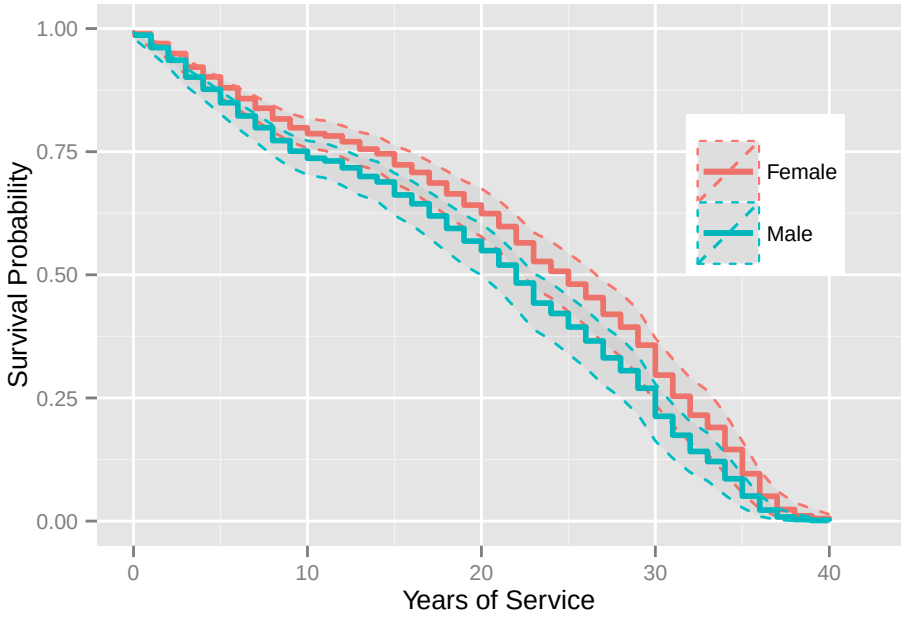


Figure 5: *Cox PHM Survival Curve, Males vs. Females.*

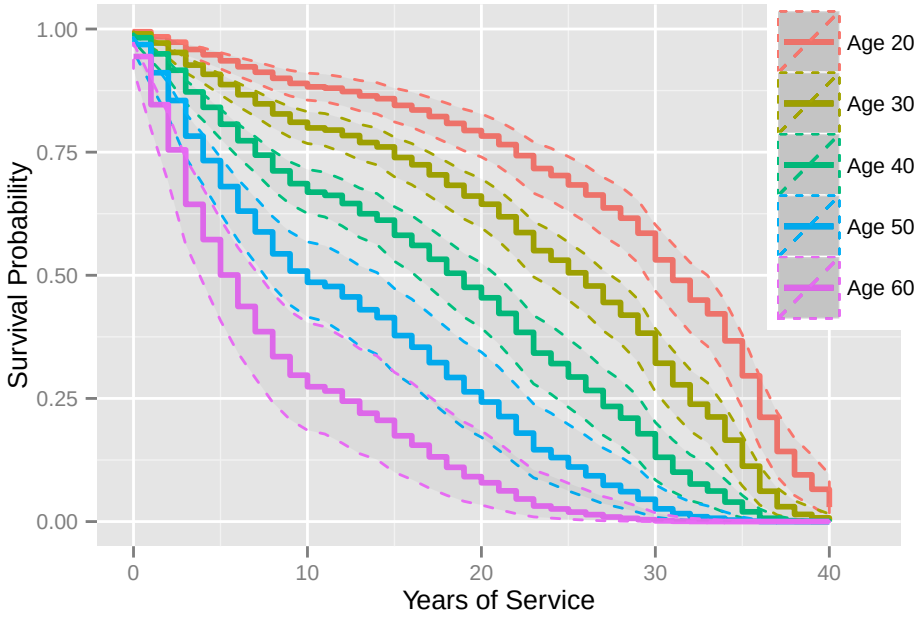


Figure 6: *Cox PHM Survival Curve, Varying Starting Ages.*

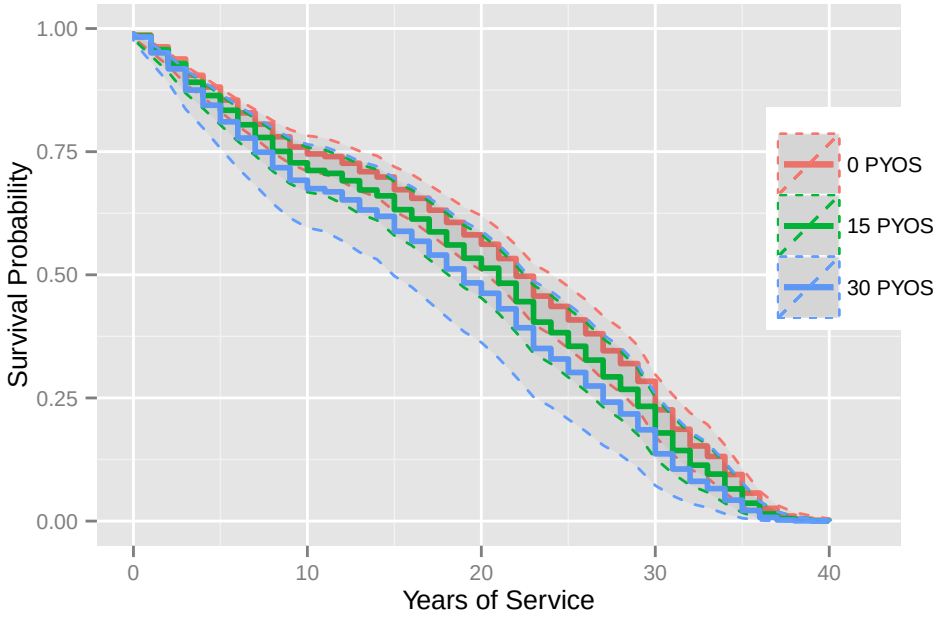


Figure 7: *Cox PHM Survival Curve, Varying Starting PYOS.*

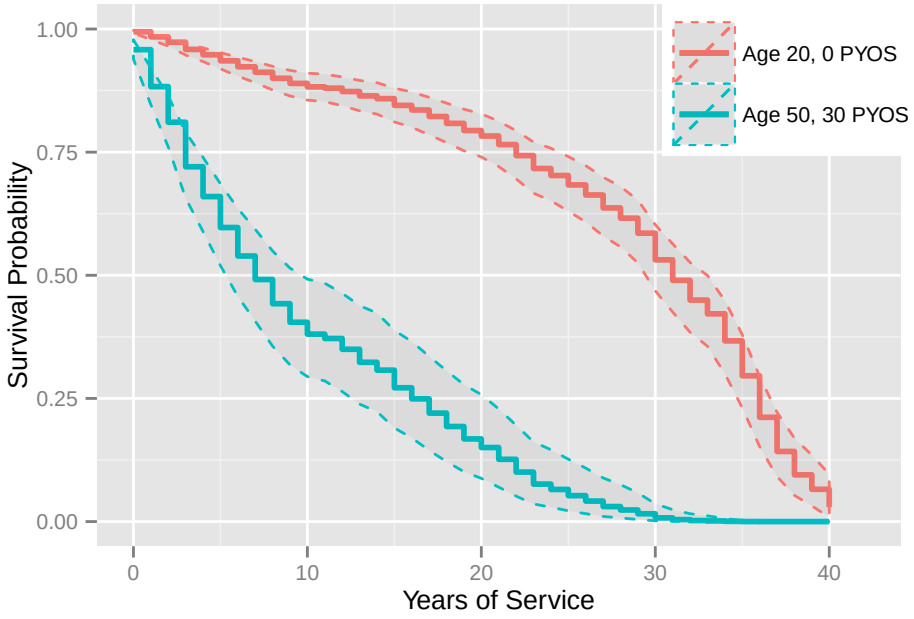


Figure 8: *Cox PHM Survival Curve, Two Extreme Cases.*

4.3 The Extended Cox Proportional Hazards Model - Complete Histories

In order for time-varying covariates to be used in the Extended Cox PHM, all individuals included in the dataset must have observations for their entire stay within the PG group. Thus, individuals who have more than 1 year of service in the beginning year of 2003 cannot be included because their history within the service would not be complete: factors may have changed in the year(s) prior to entering the dataset. Covariate information does not exist prior to this time, so starting values of important explanatory factors cannot be obtained. I therefore continue with the post-2002 group of individuals only.

The variables I employ in this section differ from those within the Cox PHM analysis in Section 4.2 in two important ways. First, the time-varying covariates 'classification level' and 'salary' are now included in the analysis. Second, I modify the attrition variable to reflect only those who leave for career-oriented reasons.[9] I therefore code all individuals who leave the PG group for other reasons such as retirement, sickness, or death, as being censored.

Employing the stepwise variable selection process described in Section 3.4, the following model is obtained:[10]

$$S(t) = \exp\left(-\int_0^t h_o(s)\exp\left[\beta_1 x_{level=2}(t) + ... + \beta_5 x_{level=6}(t) + \beta_6 x_{salary}(t) + \beta_7 x_{french}\right] ds \right) \tag{6}$$

This model thus includes 'level', 'salary', and the 'french' binary indicator. Neither 'starting age' nor 'starting PYOS' are selected according to the AIC.

Table 5 presents the regression results. Increasing in 'level' is associated with an increasingly greater proportional downward shift in the hazard function. This result is similar to those in previous analyses, such as [3] and [2]. The effect of 'salary', meanwhile, results in a lower proportional likelihood of attrition over time. There is a clear relationship between 'level' and 'salary'; a higher level always means a higher salary. However, the increase is not linear, and salaries vary quite widely even within the same level. For this reason it is justifiable to keep the two effects separate in model estimation. The model results also

[9] Specifically, I code individuals as leaving for career reasons if the 'leave reason' in the data indicates they resigned for outside employment or accepted employment at another government department.

[10] Annex B provides the test statistics for each step of the AIC stepwise variable selection process.

indicate that individuals who self-identify as 'french' speakers are more likely than others to leave for other employment opportunities.

Table 5: *Extended Cox PHM Model Results, Post-2002 Sample.*

Variable	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Level 2	1.9326	6.9073	0.4402	4.390	0.0000
Level 3	2.6909	14.7442	0.6070	4.433	0.0000
Level 4	3.1226	22.7055	0.4888	6.388	0.0000
Level 5	4.0175	55.5641	0.6308	6.369	0.0000
Level 6	5.2103	183.1430	0.7622	6.835	0.0000
Salary	-0.0734	0.9292	0.0118	-6.241	0.0000
French	0.5032	1.6540	0.1773	-2.838	0.0045

p-values are approximate. "Level" coefficients are in comparison to the effect of being at Level 1. Salary is measured in thousands.

4.3.1 Diagnostics

The Cox-Snell residuals for the extended Cox model are presented in Figure 9. Here, the residuals display only small deviations from the 45° line, indicating a reasonable, if not perfect, fit overall. Table 6 lists the results of several model diagnostics. The Gramsch-Therneau p-values indicate a failure to reject the proportional hazards assumption for each of the included factors. Further, the model fit tests all strongly reject the null hypothesis that the coefficients are equal to zero, indicating a well-specified model.

4.3.2 Comparisons

The coefficient estimates described in the previous section can be somewhat nebulous and difficult to interpret given the competing effects of level and salary. Figures 10 through 12 represent survival curves for hypothetical individuals with differing values of the covariates. The effect of the binary 'french' variable is excluded.

Figure 10 displays survival curves for individuals with levels 1 through 6. The effects of changes in level appear rather substantial; however, this comes with two caveats. First, the 95% confidence interval for the majority of these values of the covariates are overlapping. In particular, the upper end of the interval for level 6 reaches the midpoint of the level 2 confidence band. Second, for the sake of demonstrating only the effect of level and not

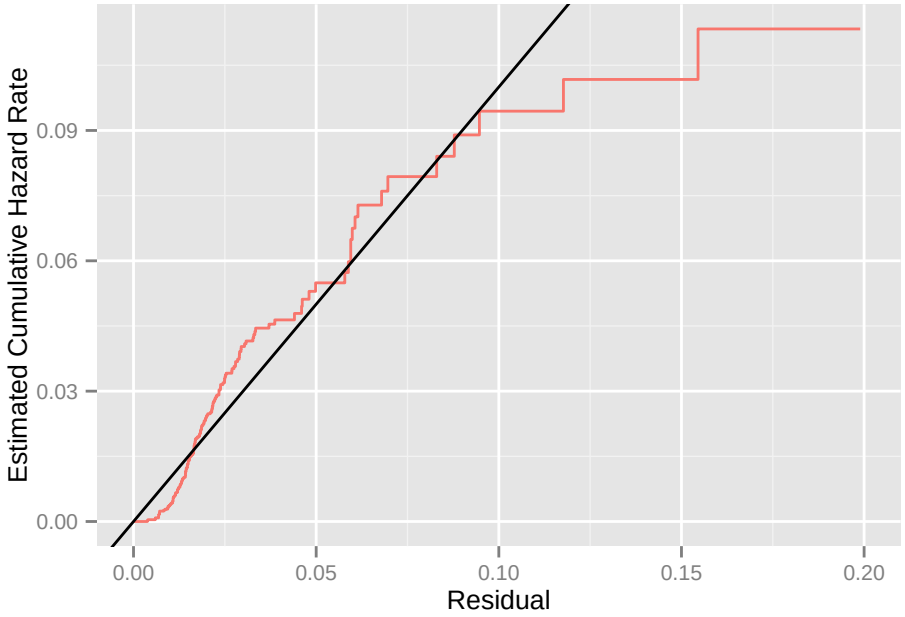


Figure 9: *Cox-Snell Residuals.*

Table 6: *Cox PHM Diagnostics, Post-2002 Sample.*

Variable	Pr(>|z|)	Grambsch-Therneau p-value
Level 2	0.0000	0.8170
Level 3	0.0000	0.4100
Level 4	0.0000	0.8620
Level 5	0.0000	0.9800
Level 6	0.0000	0.3680
Salary	0.0000	0.2970
French	0.0045	0.8170

Concordance = 0.711 (se = 0.025)
R-square = 0.11
Likelihood ratio test = 58.43 on 7 df, p = 0
Wald test = 61.88 on 7 df, p = 0
Score (logrank) test = 59.45 on 7 df, p = 0

p-values are approximate.

that of salary, a common salary - the mean of the entire sample - was set for each of these curves. This assumption is a simplifcation meant to clarify the competing effect of the level covariate, and does not correspond with usual rates of pay for each level.

Figure 11 shows the effect of salary on the underlying hazard function. I have arbitrarily chosen the 40, 60, 80 and 100 thousand dollar salary levels for this comparison. The opposing effect of salary on the attrition likelihood is quite pronounced, with lower-salaried individuals far more likely to attrite at any given moment than those with higher salaries.

To aid in the understanding of the competing effects of salary and level, I provide Figure 12 to show survival curves for hypothetical individuals at levels 1 through 6, but now assigning appropriate average salaries for each of these levels. That is, the survival curve representing an individual at level 1 is assigned a salary that is the mean for that level, and so on. This figure shows rather convincingly that while a change in level is likely to have a negative impact on the underlying survival function, the effect of salary softens this impact nearly to the point of rendering the effect statistically indistinguishable.

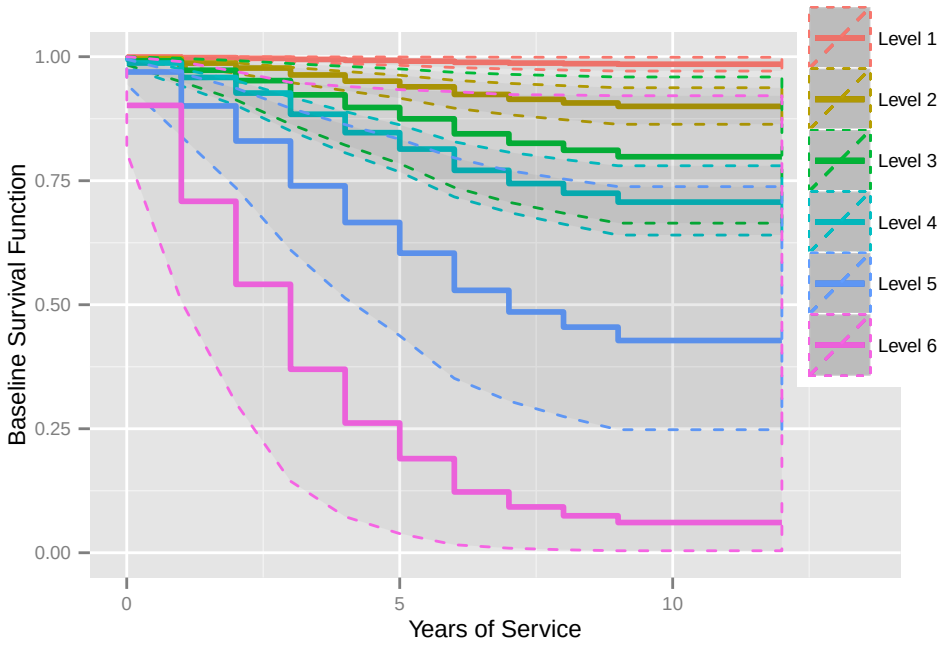


Figure 10: *Extended Cox PHM, Varying Level.*

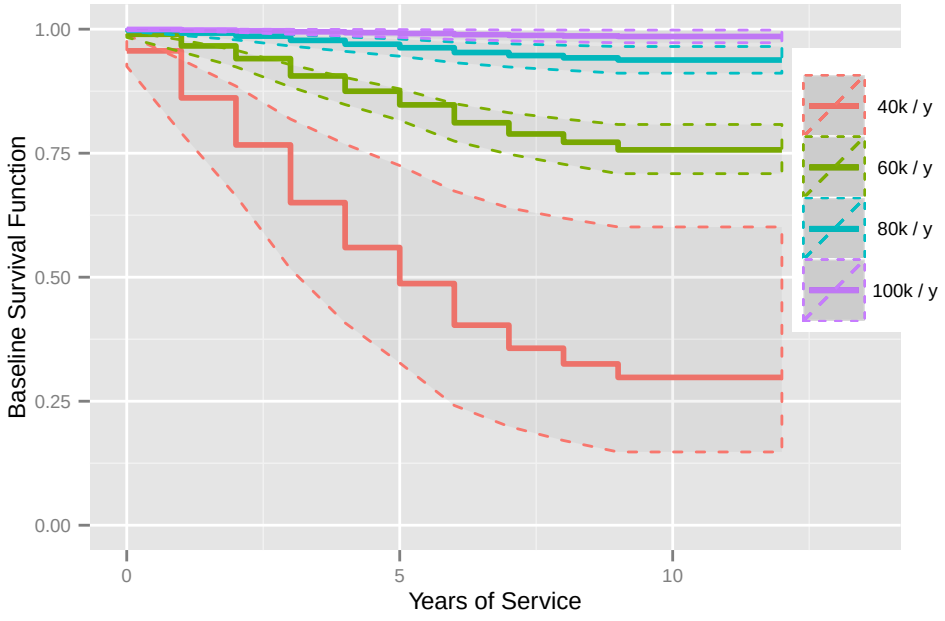


Figure 11: *Extended Cox PHM, Varying Salaries.*

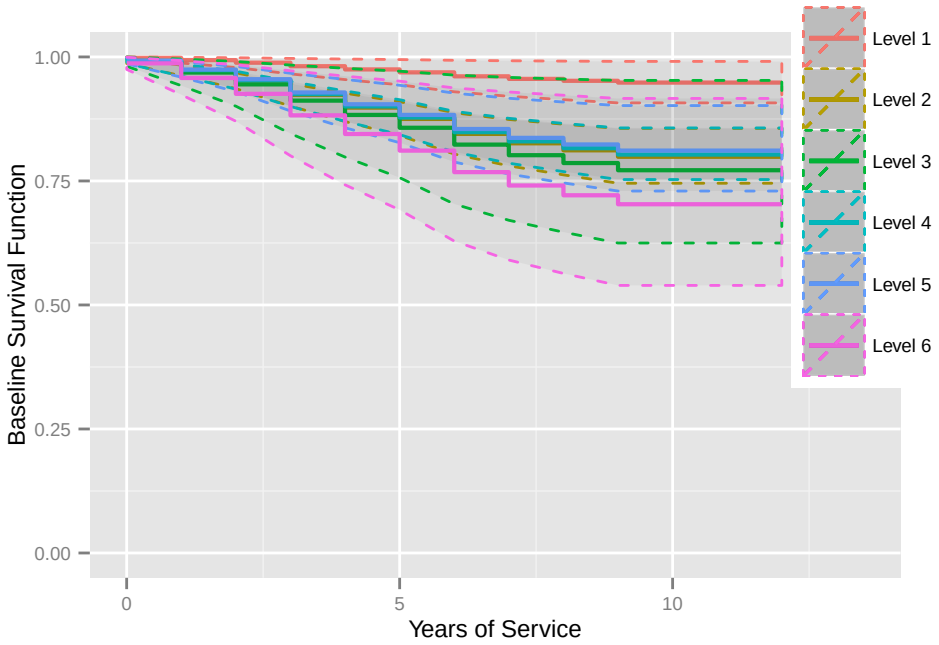


Figure 12: *Extended Cox PHM, Varying Levels while Controlling Salaries.*

5 Conclusion

This research is the result of a request by the client, Directorate Materiel Group Management Coordination (DMGMC), for an analysis of the workforce attrition phenomenon and potential economic factors therein.

According to the results of the Cox Proportional Hazards model, the problem of workforce attrition appears to mostly rely on 'starting age' and 'starting pensionable years of service'; both of these variables are shown to have a positive effect on the probability of instantaneous attrition. Differences in the attrition likelihood across genders is small, but statistically significant.

Making few epistemic assumptions and controlling only for a small number of variables, the overall phenomenon appears to be well captured, as evidenced by regression diagnostics and residual tests. These results could be used to assess expected attrition levels for future periods, and therefore better inform year-to-year projections of workforce levels.

A secondary analysis within this paper, employing the Extended Cox Proportional Hazards framework, examines the effects of financial and economic factors on attrition behaviour. The results reveal the potential existence of an effect of changes in salary and career progression (i.e., increases in classification level) on the underlying hazard, though these effects are difficult to distinguish statistically: while career progression is linked with a shorter career duration, salary is shown to decrease the instantaneous attrition likelihood.

If the client wishes further examination into the role of economic factors in workforce attrition, I recommend that they make an effort to obtain data with a far longer data horizon, extending the current sample by 10 years or more. A much larger sample of complete histories needs to be created as any financial or economic effects may be difficult to observe empirically. It would also be worthwhile to obtain data on alternate variables that also may be useful in explaining attrition behaviour, such as 'number of dependents' and 'health status'.

This page intentionally left blank.

References

[1] David W. Maybury and Christopher E. Penney. A survival analysis approach to workforce attrition data (unpublished draft). *Scientific Report, Defence Research and Development Canada*, 2015-RXX, 2015.

[2] Christopher E. Penney. Modelling voluntary attrition of civilian personnel in the department of national defence. *Contract Report, Defence Research and Development Canada*, 2013-063, 2013.

[3] Emilia Galan and Jeffrey Penney. Factors affecting public sector attrition: the case of department of national defence civilian employees - preliminary results. *Presentation at the 9th Defence and Security Economics Workshop, Carleton University*, 2015.

[4] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2005.

[5] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[6] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, B34:187–220, 1972.

[7] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[8] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

This page intentionally left blank.

Annex A Stepwise Information Criterion Results: Cox PHM

Listed below are the successive steps in the model selection process for the Cox Proportional Hazards model introduced in Section 4.2.[11]

```
Start:  AIC=6747.91
Surv(yos, attrite) ~ 1
          Df     AIC
+ st.age   1 6543.2
+ st.pyos  1 6644.9
+ female   1 6730.0
<none>       6747.9
+ french   1 6749.5

Step:  AIC=6543.2
Surv(yos, attrite) ~ st.age

          Df     AIC
+ st.pyos  1 6535.5
+ female   1 6537.5
<none>       6543.2
+ french   1 6543.3

Step:  AIC=6535.45
Surv(yos, attrite) ~ st.age + st.pyos

                 Df     AIC
+ female          1 6531.1
<none>              6535.5
+ french          1 6535.8
+ st.age:st.pyos  1 6537.4

Step:  AIC=6531.12
Surv(yos, attrite) ~ st.age + st.pyos + female

                 Df     AIC
+ st.pyos:female  1 6527.8
+ st.age:female   1 6529.8
<none>              6531.1
+ french          1 6531.2
+ st.age:st.pyos  1 6533.1

Step:  AIC=6527.84
Surv(yos, attrite) ~ st.age + st.pyos + female + st.pyos:female

                 Df     AIC
<none>              6527.8
+ french          1 6528.1
+ st.age:female   1 6529.5
+ st.age:st.pyos  1 6529.8
```

[11] The methodology for the stepwise selection process is explained in Section 3.4.

Annex B Stepwise Information Criterion Results: Extended Cox PHM

The results of the stepwise AIC information criterion approach to variable selection for the Extended Cox model given in Section 4.3 are given below.[12]

```
Start:  AIC=2026.33
Surv(yos_start, yos_end, att_employ) ~ 1

                Df     AIC
+ factor(level)  5 2017.6
+ french         1 2024.7
+ age            1 2025.4
<none>             2026.3
+ pyos           1 2026.5
+ female         1 2027.7
+ salaryk        1 2028.3

Step:  AIC=2017.59
Surv(yos_start, yos_end, att_employ) ~ factor(level)

          Df     AIC
+ salaryk  1 1987.4
+ french   1 2012.1
+ age      1 2016.7
<none>       2017.6
+ pyos     1 2018.0
+ female   1 2019.5

Step:  AIC=1987.38
Surv(yos_start, yos_end, att_employ) ~ factor(level) + salaryk

                        Df     AIC
+ french                 1 1981.9
+ age                    1 1987.2
<none>                     1987.4
+ factor(level):salaryk  5 1987.5
+ pyos                   1 1987.9
+ female                 1 1989.2

Step:  AIC=1981.9
Surv(yos_start, yos_end, att_employ) ~ factor(level) + salaryk +
    french

                        Df     AIC
+ age                    1 1981.5
<none>                     1981.9
+ factor(level):salaryk  5 1982.0
+ pyos                   1 1982.1
+ french:salaryk         1 1982.2
+ female                 1 1983.8
+ french:factor(level)   5 1988.0
```

[12] The methodology for the stepwise selection process is explained in Section 3.4.

```
Step:  AIC=1981.53
Surv(yos_start, yos_end, att_employ) ~ factor(level) + salaryk +
    french + age

                          Df     AIC
<none>                       1981.5
+ salaryk:age              1 1981.6
+ french:salaryk           1 1982.1
+ french:age               1 1982.4
+ factor(level):salaryk    5 1982.6
+ pyos                     1 1983.0
+ female                   1 1983.3
+ french:factor(level)     5 1988.2
+ factor(level):age        5 1989.6
```

Abbreviations and Acronyms

AIC Akaike Information Criterion

CORA Centre for Operational Research and Analysis

DET Defence Economics Team

DMGMC Directorate Materiel Group Management Coordination

DMGOR Directorate Materiel Group Operational Research

DRDC Defence Research and Development Canada

DSPA Directorate Strategic Planning and Accountability

PHM Proportional Hazards Model

PRI Personal Record Identification

PG Purchasing and Supply Group

PYOS Pensionable Years of Service

DOCUMENT CONTROL DATA

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g., Centre sponsoring a contractor's report, or tasking agency, are entered in Section 8.)

DRDC – Centre for Operational Research and Analysis
Dept. of National Defence, MGen G. R. Pearkes Bldg., 101 Colonel By Drive, Ottawa
ON K1A 0K2, Canada

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

UNCLASSIFIED

2b. CONTROLLED GOODS

(NON-CONTROLLED GOODS)
DMC A
REVIEW: GCEC APRIL 2011

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

A survival analysis of workforce attrition : Examining attrition phenomenon through non-parametric and semi-parametric survival methdos

4. AUTHORS (last name, followed by initials – ranks, titles, etc., not to be used)

Christopher E. Penneythor]

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
October 2016	26	8

7. DESCRIPTIVE NOTES (The category of the document, e.g., technical report, technical note or memorandum. If appropriate, enter the type of report, e.g., interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Scientific Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC – Centre for Operational Research and Analysis
Dept. of National Defence, MGen G. R. Pearkes Bldg., 101 Colonel By Drive, Ottawa
ON K1A 0K2, Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC-RDDC-2016-R162	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

UNLIMITED

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

UNLIMITED

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

This paper explores the problem of workforce attrition and the role of economic factors in attrition behaviour. There are two main elements to this study: first, I examine the overall attrition phenomenon using non-parametric and semi-parametric models; second, I consider the role of economic factors by limiting the sample to individuals for whom complete histories are available and then employing a semi-parametric model that allows for time-varying elements.

The findings for the primary analysis indicate a strong model fit can be obtained through the inclusion of a few explanatory variables: these are 'starting age', 'starting pension-able years of service', and 'gender'. The findings for the secondary analysis provide some evidence of a role in attrition behaviour for economic factors such as 'salary' and 'classi-fication level', but the effects are offsetting and difficult to distinguish statistically.

I recommend that the client provide data with a significantly longer time horizon so that a more precise determination of the role of economic and financial factors can be obtained.

Dans le présent document, j'explore le problème de l'attrition des effectifs ainsi que le rôle des facteurs économiques dans la tendance en matière d'attrition. L'étude comporte deux grands volets : dans un premier temps, j'examine le phénomène de l'attrition dans son ensemble au moyen de modèles non paramétriques et semi paramétriques ; dans un deuxième temps, j'étudie le rôle des facteurs économiques en utilisant un échantillon com-posé uniquement de personnes dont l'historique complet est accessible, puis en employant un modèle semi paramétrique permettant la prise en compte d'éléments variant dans le temps. Les résultats de l'analyse initiale indiquent qu'il est possible d'obtenir un modèle bien ajusté grâce à l'inclusion de quelques variables explicatives, à savoir : l'âge au début, le nombre d'années de service ouvrant droit à pension au début et le sexe. Les résultats de la seconde analyse révèlent que certains facteurs économiques comme le salaire et le niveau de classification jouent un rôle dans la tendance en matière d'attrition, mais que leurs effets sont conflictuels et difficiles à cerner statistiquement parlant. Je recommande au client de fournir des données avec un horizon temporel beaucoup plus long de façon à pouvoir faire une évaluation plus précise du rôle des facteurs financiers et économiques.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g., Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Workforce attrition, survival analysis, personnel modelling, attrition risk

DRDC | RDDC

SCIENCE, TECHNOLOGY AND KNOWLEDGE FOR CANADA'S DEFENCE AND SECURITY | SCIENCE, TECHNOLOGIE ET SAVOIR POUR LA DÉFENSE ET LA SÉCURITÉ DU CANADA

www.drdc-rddc.gc.ca