8-2016

Forecasting Employee Turnover in Large Organizations

Xiaojuan Zhu
University of Tennessee, Knoxville, xzhu8@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaojuan Zhu entitled "Forecasting Employee Turnover in Large Organizations." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Rapinder Sawhney, Major Professor

We have read this dissertation and recommend its acceptance:

John E. Kobza, James L. Simonton, Russell Zaretzki

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Forecasting Employee Turnover in Large Organizations

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Xiaojuan Zhu

August 2016

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Rapinder Sawhney, for his guidance, caring and patience during the course of my academic studies, research and dissertation. I would like to thank my committee members, Dr. Russell Zaretzki, Dr. John E. Kobza, and Dr. James L. Simonton for their help and guidance throughout my research.

Finally, I would like to express thanks to the past and current members of Dr. Sawhney's group for their help during my research. I also appreciate all the assistance and insights provided by Dr. William Seaver, Dr. Shuguang Ji and Dr. Bruce A. Holt.

Abstract

Researchers and human resource departments have focused on employee turnover for decades. This study developed a methodology forecasting employee turnover at organizational and departmental levels to shorten lead time for hiring employees. Various time series modeling techniques were used to identify optimal models for effective employee-turnover prediction based on a large U.S organization's 11-year monthly turnover data. A dynamic regression model with additive trend, seasonality, interventions, and a very important economic indicator efficiently predicted turnover. Another turnover model predicted both retirement and quitting, including early retirement incentives, demographics, and external economic indicators using the Cox proportional hazard model. A variety of biases in employee-turnover databases along with modeling strategies and factors were discussed. A simulation demonstrated sampling biases' potential impact on predictions. A key factor in the retirement was achieving full vesting, but employees who did not retire immediately maintain a reduced hazard after qualifying for retirement. Also, the model showed that external economic indicators related to S&P 500 real earnings were beneficial in predicting retirement while dividends were most associated with quitting behavior. The third model examined voluntary turnover factors using logistic regression and forecasted employee tenure using a decision tree for four research and development departments. Company job title, gender, ethnicity, age and years of service affected voluntary turnover behavior. However, employees with higher salaries and more work experience were more likely to quit than those with lower salaries and less experience. The result also showed that college major and education level were not associated with R&D employees' decision to quit.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Employee turnover refers to an employee's leaving his/her job as a result of voluntary quitting, retirement, disability, or death. Employee turnover has drawn management researchers' and practitioners' attention for decades because turnover cost affects an organization's operational capabilities and budget. Employee turnover is both costly and disruptive to the organizational function (Kacmar et al., 2006; Mueller and Price, 1989; Staw, 1980); and both private firms and governments spend billions of dollars every year managing this issue (Leonard, 2001). Turnover costs involve recruiting, selecting, and training (Mobley, 1982; Staw, 1980). According to the U.S. Department of Labor, turnover costs a company one-third of a new hire's annual salary to replace an employee, which is about $500 to $1500 per person for the fast-food industry and $3000 to $5000 per person for the trucking industry (White, 1995).

Furthermore, turnover disrupts social and communication structures and causes productivity loss (Mobley, 1982). Turnover also demoralizes the remaining employees and leads to additional turnover (Staw, 1980). Sagie et al. (2002) found that a high-tech firm lost 2.8 million US dollars or 16.5% of before-tax annual income because of employee turnover. These researchers also found that turnover reduced profits, increased the organization's total risk, and triggered more turnover among the organization's other employees. Therefore, understanding and forecasting turnover at the firm and departmental levels is essential for

reducing it (Kacmar et al., 2006), as well as for effectively planning, budgeting, and recruiting in the human resource field.

Many studies have shown that employee turnover significantly affects an organization's performance. Staw (1980) summarized previous studies indicating that turnover reduces organizations' team or work performance and financial performance and significantly changes an organization's direction when a top executive leaves. Glebbeek and Bax (2004) found that excessive employee turnover harms firm performance (profits). According to Hancock et al. (2013)'s study, turnover has a strong negative relationship with organizational performance in the manufacturing and transportation industries. According to Kacmar et al. (2006)'s study, employee turnover increases customer waiting time and reduces profits. Turnover also reduces restaurants' profitability and customer satisfaction because of declining productivity (Detert et al., 2007). On the other hand, the lower turnover rate increases sales (Batt, 2002). Overall, the inability to predict employee turnover and to replace that individual reduces organizational performance and profits and disrupts the organizational structure.

The value of being able to predict turnover is directly related to the time. It takes an organization to hire and on-board employees as compared to their needs. This time is not only dependent on the organization, but possibly the type of employees that are being hired. Production-based governmental organizations that are the basis of this analysis require a lengthy period to hire and on-board employees, because of clearance and security issues. Predicting turnover early allows HR to proactively plan for the possible turnover and to be prepared to launch the search, shortening the time-to-hire and on-board employees. Therefore, Predicting employee turnover helps reduce the hiring lead time, thus eliminating some turnover costs. The lead time of employee turnover and replacement involves five stages: providing a leaving notice, advertising the job opening, interviewing, doing a background check, and completing on-board training. An employee usually provides a leaving notice at least two weeks before the actual leaving date. Then the employee's manager must prepare for advertisement and hiring committees, requiring a week to complete. The hiring committees need two to eight weeks, sometimes even longer, to release the hiring advertisement and select interview candidates. During the interview period, hiring committees need one to two weeks to make a final decision. One week to six months is needed to complete the finalist's

background security check. Finally, the employee's on-board training takes at least one week. Thus, an organization usually spends at least two months replacing a new employee. For some governmental organizations, the hiring lead time is much longer since their security check takes three to six months. The forecasting system provides the predicted turnover number a year in advance. Based on a combination of this number, the production plan, and the budget, the HR department determines a final demand number. Ignoring employees' notice periods, HR can either advertise job openings and start hiring immediately based on the final demand number (thus reducing the training period) or wait until an employee provides the leaving notice. Therefore, forecasting turnover at the firm and departmental levels reduces hiring lead time (Kacmar et al., 2006). The other benefit is to identify unusual patterns in turnover and possibly investigate root causes.

1.2 Purpose of Study

Some organizations have a long hiring lead time because of the time required for security background checks ($>$ 6 months). This study developed a methodology to forecast employee turnover at organizational and departmental levels to shorten this time. It also investigated turnover seasonal patterns and such factors as employee demographics, job categories, and organization structures; identified influential financial indices; and measured the magnitude of retention and the Early Retirement Incentive Program (ERIP) (Clark, 2002) that HR released; and forecasted employees' turnover based on retirement and voluntary quitting. The study also simulated turnover datasets to measure two data biases and examines the forecasting models' forecasting capability. Finally, the study examined job title, gender, ethnicity, age, and years of service to assist in employee-retention strategies to reduce R&D departments' voluntary turnover rate.

1.3 Approach

Forecasting employee turnover is a crucial part of a lean management system (Allway and Corbett, 2002). Also called the Toyota production system, lean management involves

operating the most efficient and effective organization possible with the least cost and *zero* wastes (Jackson, 1996). The waste is non-value-added activities. High employee turnover is one kind of waste (Kilpatrick, 2003). According to Allway and Corbett (2002), employee turnover is a significant factor in lean management.

An employee-turnover forecasting system reduces wastes that high turnover causes. By identifying factors influencing employee turnover, the forecasting system assists HR in determining retention strategies to reduce turnover rates. According to Yeung and Berman (1997) and Kochan and Useem (1992), "high-commitment" human resource polices contribute to a lean management system's success.

Furthermore, an employee-turnover forecasting system reduces the setup time of the organization's lean-management system. According to Lin and Hui (1999), employee turnover takes a lean organization much more time and far more resources. Productivity is usually reduced because an employee leaves and the replacement is unfamiliar with the work. The new employee needs a training period to be as productive as the previous employee. To prevent productivity reduction, HR could hire a new replacement based on the demand number computed from employee-turnover forecasting. Then a new employee could receive job training before the old employee leaves the position. As a result, productivity level could remain the same before and after employee turnover.

Finally, an employee-turnover forecasting system periodically provides current and future inventory levels. The inventory management system uses the demand number the system provides, the hiring lead time, and the hiring costs to determine the optimal hiring number through an appropriate economic order model. As a result, the skill-set inventory is systematically controlled.

1.4 Methodology

Statistical methods are used to develop employee-turnover forecasting models. These methods are time series, survival analysis, logistic regression, and data mining. Time series methods capture employee turnover's seasonal and cyclical patterns and forecast an aggregated turnover number, in terms of headcount, by using a historical turnover number.

Survival analysis identifies significant internal and external turnover factors and builds a Cox PH model to forecast turnover at the individual, departmental, and entity-wide levels. Logistic regression method determines whether significant factors identified in the literature are also significant in the organization studied. A set of decision rules is created based on employees' tenure using a decision tree method.

Implementing an employee-turnover forecast model is also a key part of workforce planning in a lean management system. The model is inserted into a software program using a user-friendly interface. The human resource department installs this program, imports the employee's information into the program periodically, computes forecasting information, and exports the results. Based on the results, HR either modifies the employee retention and promotion strategies for the targeted employees with high turnover probabilities to reduce the employee turnover rate or prepares to hire new employees to prevent reduced productivity and lean-management system malfunction.

1.5 Outline

This dissertation includes five chapters. Chapter 1 introduces employee-turnover forecasting. Chapter 2 introduces time series methods and identifies an optimal time series model to forecast the monthly employee-turnover number. Chapter 3 discusses how to build a Cox proportional model to identify the significant demographics, organizational factors, and financial indices as well as to forecast employee retirement and voluntary quitting on an individual level. Chapter 4 presents the hypotheses regarding employee turnover significant factors for an R&D organization and splitting rules used to forecast the average tenure. Chapter 5 summarizes the findings, the implementation of forecasting model, and future work.

Chapter 2

Employee Turnover Forecasting for Human Resource Management Based on Time Series Analysis

2.1 Introduction

Employee turnover has drawn researchers' and human resource managers' attention because organizations lack niche skill sets and resources, which require time and planning to acquire at crucial times. The hiring lead time is often long, particularly when special skills are involved. In some organizations, like U.S. national laboratories, the process can take months because of security-clearance requirements. Therefore, a good employee-turnover prediction at the firm and departmental levels is essential for effective human resource planning (HRP), budgeting, and recruiting.

Human resource planning is an ongoing systematic planning process to optimize the human resource pool. For organizations to efficiently and effectively execute tasks, the right people must be available at the right places at the right time (Khoong, 1996). Over the years, organizations have scaled up their efforts in manufacturing, marketing and financing. However, organizations have always struggled to develop sustainable HRP models (Heneman et al., 1993), whose objective is to match employees and jobs to avoid manpower shortages or

surpluses (Čambál et al., 2011). To achieve this balance, employee turnover is often central to organizational workforce planning and strategy.

As summarized in Table 2.2, researchers developed employee turnover models using various statistical methods. Previous studies have identified employee-turnover explanatory predictors. For instance, Bluedorn (1982) related turnover to the individual's routine, age, service length, and perception of environmental opportunities. Balfour and Neff (1993) suggested that caseworkers with more education, less experience, and less stake in an organization are more likely to turnover. Wright and Cropanzano (1998) associated emotional exhaustion with job performance and subsequent turnover, but not with job satisfaction. Predicting employee turnover based on employee absenteeism and performance, Morrow et al. (1999)'s study showed a positively correlated absenteeism and voluntary turnover as well as negatively correlated performance ratings and voluntary turnover. Thaden et al. (2010) indicated that organizational culture might potentially be an important factor for retaining workers.

Other insights have been gained from more recent research. For instance, according to Tews et al. (2014), personal events, professional events, internal work events and constituent attachment are highly related to turnover. Collini et al. (2015) found that the interaction between interpersonal respect, mission fulfillment, and engagement are statistically significant predictors for turnover in health care. However, these researchers found that diversity climate is not related to turnover. Finally, only Sexton et al. (2005) considered outside economic variables, unemployment index, and consumer price index in the employee-turnover forecasting model. However, their final model did not include these variables. Ferrara and van Dijk (2014) in the *International Journal of Forecasting* revealed a new interest in forecasting business cycles with some complex methodologies. However, forecasting business-cycle turning points is quite difficult, and Hamilton (2011) suggested that "the best econometricians can do is probably to nowcast recessions; that is, to recognize a turning point as soon as it occurs, or soon thereafter." An outside variable might facilitate this situation.

Table 2.1: Summary of Previous Research on Employee Turnover Forecast

Authors (Year)	Data Acquisition	Data Horizon	Methods	Software	Economic Indicator	Response Variable	Estimate	Model Evaluation
Bluedorn (1982)	Employee records and Survey	1 year	Correlations, multiple regression	N/A	No	Number	Point with intervals	$R^2 = 0.22$, Adjusted $R^2 = 0.11$
Ng et al. (1991)	Survey	N/A	Hazard proportional model	BMDP 2L	No	Probability	Point with intervals	Pair t-test
Balfour et al. (1993)	Employee records	33 months	Non-linear logistic regression	N/A	No	Probability	Point	Chi-square values
Feeley et al. (1997)	Survey	60 months	Social network, logistic regression, correlation	NE-GOPY, UCINET	No	Probability	Point	$R^2 = 0.23$
Wright et al. (1998)	Survey	1 year	Hypothesis test, correlation, logistic regression	N/A	No	N/A	N/A	Correlation r=0.34, $P < 0.01$
Morrow et al. (1999)	Demographic information and employee records	2 years	Logistic regression, correlation	N/A	No	Probability	Point	(-2 log likelihood) chi-square=193.13
Sexton et al. (2005)	Demographic information and employee records	10 years (yearly)	NN	FOR-TRAN	Yes	Leave or not	Point	Type I error=0.25% Type II error=5.83%
Hong et al. (2007)	Survey	N/A	Logit and probit model	SPSS	No	Probability	N/A	$R^2 = 0.5$, Quadratic Probability Scores = 0.18 for training and 0.12 for test
Nagadevara et al. (2008)	Demographic information and employee records	3 years	NN, logistic regression, classification/regression trees, discriminant analysis	N/A	No	Leave or not	Point	Contingency table
Thaden et al. (2010)	Survey	2 years	Multiple regression	N/A	No	Duration	Point with intervals	$R^2 = 0.56$, $P < 0.001$
Grler and Zock (2010)	Employee records	360 months	System dynamics	N/A	No	Number	Point	N/A
Saradhi et al. (2011)	Survey	2 years	SVMs, random forest, Nave Bayes classifiers	N/A	No	Probability	Point	True/false positive rate and precision
Alao et al. (2013)	Employee records	28 years (yearly)	Decision tree	WEKA See5	No	Probability	Point	True/false positive rate and precision
Tews et al. (2014)	Employee records and Survey	6 months	Logistic regression	N/A	No	Probability	Point	$R^2 = 0.23$
Collini et al. (2015)	Survey and turnover rates	1 year	Correlation and Regression	N/A	No	Turnover rates	Point	No

Meanwhile, other studies have tried to build turnover prediction models through such techniques as regression, neural network (NN), and data mining. For example, Ng et al. (1991) used a proportional hazards regression (PHR) to develop a turnover-prediction model. In Sexton et al. (2005)'s study, NN combined with a modified genetic algorithm was used to build a turnover-prediction model. Alao and Adeyemo (2013) applied a decision tree to the employees' demographical information and personnel records to identify attributes contributing to employee turnover. In these studies, the data source was acquired from either human resource employment records and demographical information or surveys with time horizons ranging from 1 to 28 years. Most of the data were monthly, which is ideal for time series forecasting models.

Although some efforts have been made to predict employee-turnover behavior, no study has investigated the prediction of employee turnover with time series forecasting techniques. Therefore, this study attempted to fill this research gap. The advantages of a time series forecasting approach are that identifying turnover's determinants is unnecessary and evaluating either a planned or unplanned intervention's effects is helpful (Velicer and Fava, 2003).

This chapter has four parts. The introduction covers the paper's objective and a literature review. Part two identifies tools used in finding time series patterns and in preparing the data for analysis as well as specific forecasting methods. Part three provides the study's results. Part four includes the study's practical implications and limitations.

2.2 Methods

2.2.1 Data Preparation

A large multipurpose research organization in the U.S. provided the human resource data. The dataset consisted of over 8,000 observations of active and terminated employees' with incomplete demographic information (including metrics such as payroll category, hired date, termination date, age, years of service, gender, job classification, and department code) from November 2000 to January 2012. The turnover dataset was summarized in the

form of monthly data in which each field represented the number of employees leaving the organization. For this research, turnover is defined as the total number of employees leaving the organization each month. This definition is used as a unit of measurement for the turnover prediction.

This study examined several economic indicators: unemployment rate index, New York Stock Exchange, U.S gasoline price, and U.S monthly composite leading indicator (CLI). However, only the CLI that the Organization for Economic Co-operation and Development (OECD) published from November 2000 to January 2012 significantly improved the forecasting performance as a predictor of employee turnover's cyclical component because, in practice, the CLI is an early indication of turning points in the macroeconomic cycle. OECD constructed the CLI data by aggregating seven components: the number of dwellings started, net new orders for durable goods, share prices-NYSE composite, consumer-sentiment indicator, average weekly hours manufacturing workers worked, purchasing manager index, and spread of interest rates (Organisation for Economic Co-operation and Development, 2013). The index construction's weights are not released by OECD.

2.2.2 Pattern Analysis and Outlier Identification

To model a time series, looking for patterns in the turnover series is important. First, it is simply applied to a time plot of the series and box plots of the seasons or months. In this case, the turnover series' seasonal pattern was tested using Kruskal-Wallis and ANOVA tests ($P < 0.05$), which did not correct for any trend in the series. The second stage of the pattern analysis involves autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify seasonal, autoregressive and moving average patterns. If external variables exist (as in this case), the third stage of pattern analysis examines the cross-correlations between turnover series (Y_t) and external variables (CLI (X_t) over time). The cross-correlation function (CCF) was used to identify lags of CLI(X_t)), which might be useful predictors of the turnover series (Y_t). A longer lag that is strong enables the forecast horizon to be longer when using an external variable.

All of the previous stages of the pattern analysis can be contaminated by outliers, so identifying outliers before fitting the actual forecasting model is important. Box plot

analysis by seasons can be used to informally flag outliers; however, this approach tends to over-identify outliers. In contrast, ARIMA methods in conjunction with statistical process control (SPC) tend to correctly identify the number of outliers in a series. Conservative ARIMA models ((1,0,1) or (0,1,1)) are used in this control charting. First, the residuals from ARIMA (1,0,1) are divided by the root mean square error to standardize them, and then the outliers are identified with the value greater than ±3 standard deviations from zero by the standardized residuals' scatterplots (Alwan and Roberts, 1988; Grznar et al., 1997). Identifying the outliers using SPC and then smoothing them is a way to both refine the data for further analysis and facilitate finding the underlying pattern in the data series. Smoothing outliers in a time series is essential to eliminate random noise and other irregularities. When the outlier is identified, it is adjusted to be more similar to its neighboring points (Grznar et al., 1997). In this study, unusual observations were smoothed using a nonlinear data smoothing method, based on repeated medians (RMD) of a five-period span (as shown in Equation 2.1) (Velleman, 1980),

$$S_t = Median(y_{t-2s}, y_{t-s}, y_t, y_{t+s}, y_{t+2s}) \tag{2.1}$$

where S_t is the actual smoothed value at time t, y_t is the value of response variable at time t, and s is the number of total seasonal periods, which is 12. This smoothing is only performed on potential outliers within their season and not across adjacent periods. Sometimes outliers can distort normality, white noise, cross-correlations, ACF and PACF, and the model's predictive performance. Thus, properly dealing with outliers means asking hard questions about these unusual observations from a human resource perspective; in turn, answering those questions can add understanding and improved forecasting ability. The outliers are not always unusual events but interventions or change points needing to be accommodated in the model. In this kind of human-resource dataset, such abnormalities could be the following: retirement incentives were offered, another company was purchased, a section of the original company was sold, or the potential outlier might reflect an economic downturn. Many possibilities of abnormalities exist; but if outliers are not over identified, a good human resource department should be able to provide quick answers in unusual situations.

2.2.3 Time Series Analysis

In time series forecasting, past observations of the variable are collected and analyzed to develop a model describing the pattern. This modeling approach is particularly useful when little knowledge is available regarding the data-generating process or when no satisfactory explanatory model relates the prediction variable to other explanatory variables (Zhang, 2003). In this study, univariate and multivariate time series methods were used to identify an optimum forecast. Number Cruncher Statistical System (NCSS), SAS, and R were the statistical software used. Two data partitions were analyzed: the training sample (November 2000January 2011) and the holdout sample (February 2011January 2012). For each model, the training sample was used to build the model while the holdout sample was used to validate the model because the most recent time series data is considered the most important factor for prediction purposes (Bergmeir and Benítez, 2012).

2.2.4 Univariate Methods (without External Variables)

Univariate time series analysis models without external variables use months for seasonality and trend as turnover predictors. The univariate models used in this study were time series regression, decomposition, Winter's Exponential Smoothing (WES), and Box-Jenkins Autoregressive Integrated Moving Averages (ARIMA). One critical reason for selecting these models was to avoid restricting the forecasting horizon, i.e., how far in the future predictions could be made.

2.2.4.1 Time Series Regression

The univariate time series regression model used a trend and seasonality as predictors. The additive time series regression model with intercept, trend, monthly seasonality, and error terms took the form shown in Equation 2.2,

$$Y_t = \beta_0 + \beta_1 x_{trend} + \sum_{i=2}^{12} \beta_i d_i + \xi_i \tag{2.2}$$

where, Y_t is response variable at time t, β_i is the coefficients estimated by regression, x_{trend} is a continuous variable representing trend with values from 1 to n, d_i is a dummy variable representing seasonal periods. In addition, the additive regression models with interventions (pulse or steps) were also analyzed because of the downsize policy in certain time points, which are denoted by a dummy variable in the regression function. The multiplicative time series regression model with intercept, trend, monthly seasonality, and error terms was also considered with the form as shown in Equation 2.3,

$$Ln(Y_t) = \beta_0 + \beta_1 x_{trend} + \sum_{i=2}^{12} \beta_i d_i + \xi_i \tag{2.3}$$

where, $Ln(Y_t)$ is natural log transformation of Y_t. For these regression models, the significance of the model and the variables were examined by using p-value at 0.05 significance level, lack of collinearity, good holdout performance, and valid regression assumptions.

2.2.4.2 Decomposition

Decomposition time series methods attempt to separate the series into four components: trend, cycle, seasonality, and irregularity. Multiplicative decomposition methods are expressed globally in Equation 2.4.

$$Y_t = f(trend, cycle, seasonality, irregularity) = T_t \cdot C_t \cdot S_t \cdot I_t \tag{2.4}$$

The decomposition model can be used assuming no cyclical variation exists, or the cyclical variation can be extracted and fit a model to enhance forecasting. Decomposition models can be quite complex, but the classical multiplicative decomposition model in NCSS was used for this turnover series as an easier application.

2.2.4.3 Winters Exponential Smoothing (WES)

WES models work well on series that have either seasonality or both seasonality and trend. The models can also be additive or multiplicative, but the preferred option tends to have an additive trend and either additive or multiplicative seasonality. The multiplicative trend

in these kinds of time series models tends to over or under forecast future values. Although dampened models can help avoid this future forecast problem, care must be taken because the forecasts will have a short horizon (DeLurgio, 1998). The robustness and accuracy of Winter's exponential smoothing methods have led to their widespread use in applications where a large number of series necessitates an automated procedure (Winters, 1960; Taylor, 2003). WES model is easy to interpret and easily understood by management or those less technically inclined.

2.2.4.4 Autoregressive Integrated Moving Average (ARIMA)

Statisticians George Box and Gwilym Jenkins introduced the ARIMA models (Box et al., 1970), whose general form is ARIMA (p,d,q) where p, d, and q are non-negative integers that refer to the order of the model's autoregressive, differencing, and moving average parts, respectively. In addition, ARIMA models can handle seasonality, in which case their forms are ARIMA (p,d,q)(P,D,Q). Thus, a series' seasonal factor could have autoregressive, differencing, or moving average patterns. Because they provide a wide class of models for univariate time series forecasting, ARIMA methods are popular with some forecasters (Harvey and Todd, 1983).

2.2.5 Univariate Methods (with External Variables)

Univariate models incorporating an external variable (CLI) as a predictor of a turnover series' cyclical component were also examined. It included dynamic regression and more complex decomposition models.

2.2.5.1 Dynamic Regression with External Variable

The dynamic regression model describes how the forecasting output is linearly related to current and past values of one or more input series. Two assumptions are critical for the dynamic regression model. First, the input series' observations occur at equally spaced time intervals. Second, the output does not affect the input series (Pankratz, 2012). Dynamic regression models allow inclusion of external variables, interventions, and transfer functions.

In this study, the external variables (CLI and interventions) were incorporated into the dynamic regression model. Equation 2.5 is a simple representation of the model with trend, seasonal dummy variable, and interventions:

$$Y_t = \beta_0 + \beta_1 x_{trend} + \sum_{i=2}^{12} \beta_i d_i + \varpi_0 I_t + \frac{\varpi_1}{1 - \delta_1 B} I_t^{'} + \xi_i \tag{2.5}$$

where I_t is a dummy variable representing pulse and step periods, and $I_t^{'}$ is a dummy variable representing pulse periods. Also, ϖ_0, ϖ_1, and θ_1 are the change-point coefficients that regression estimates and $1 - \delta_1 B$ refers to the delayed rise or fall in the forecast variable.

2.2.5.2 Decomposition with External Variable

In this more complex decomposition model, CLI and its lag terms predict a cyclical component in the decomposition model. This research applied two approaches to obtain a decomposition model: the automatic decomposition in NCSS with the cyclical variable (CLI) incorporated, and a multiplicative decomposition model built with the product of the best ARIMA and cyclical factor (CLI).

2.2.6 Nonlinear and Multivariate Methods

2.2.6.1 State Space Model

A state space model consists of an observation equation as shown in equation 2.6 and a Markovian transition equation as shown in Equation 2.7

$$y_t = F_t \theta_t + v_t \tag{2.6}$$

$$\theta_t = G_t \theta_{t-1} + w_t \tag{2.7}$$

where y_t is a $m \times 1$ data vector; θ_t is a $p \times 1$ unknown state-vector; F_t is a $m \times p$ state vector relating the observation data to the state vector θ_t; and G_t is a $p \times p$ state transition matrix. In addition, v_t, w_t are random error matrices which independently and identically

follow a multinomial distribution. State space models were modeled using R `stats`, `dlm`, and `forecast` packages.

2.2.6.2 Vector Autoregressive Model

Vector autoregressive model (VAR) is an econometric model for multivariate time series analysis. It is an extension of the univariate autoregressive model, i.e., each variable is represented as a linear function of its lags and the other variables' lags. VAR models are often used to describe and forecast financial and economic time series. A VAR model consists of a set of k variables (also called endogenous variables) $y_t = (y_{1t}, y_{2t}, \ldots, y_{kt})$, denoted as $k \times 1$ vector. A p^{th} order VAR model is expressed as

$$y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + e_t \tag{2.8}$$

where $A_i(k \times k)$ is a coefficient matrix. Also, e_t is a $k \times 1$ unobservable white noise vector process where expected value of this vector is zero and a time-invariant covariance matrix $E(e_t e_t^{'}) = \sum$. The R `vars` package models the VAR model.

2.2.7 Model Evaluation

A good forecasting model should be evaluated in term of predictive ability, goodness of fit using the R^2, mean absolute percentage error (MAPE), mean absolute error (MAE) or other fit diagnostics, normality tests on residuals, and a white noise test on those same residuals to ensure that no time series pattern is left. The best fitting model is generally selected based on a higher R^2 value in the holdout data, a lower MAPE, normally distributed residuals, and passing a white noise test.

To evaluate the time series methods, the pseudo R^2 for training and holdout data were calculated as standard criteria to test the goodness of model fitting as expressed in Equation 2.9.

$$R^2_{pseudo} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)}{\sum_{t=1}^{n}(y_t - \bar{y}_t)} \tag{2.9}$$

MAPE was another measure of the time series model fitting methods' accuracy as shown in equation 2.10 (Hanke et al., 1998; Bowerman et al., 2005). This measure was used to compare the model performance for the specific dataset by using time series methods since it measured relative performance (Chu, 1998) as reflected in Equation 2.10.

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|}{n} \tag{2.10}$$

A good time series model should have normally distributed residuals. In this study, the residuals' normality was evaluated using two powerful normality tests: Shapiro-Wilk (Shapiro and Wilk, 1965) and D'Agostino Omnibus normality test (D'agostino et al., 1990) in NCSS.

The white noise test is performed on the residuals to check for any undetected time series pattern still remaining in the residuals and not accounted for by the model. Independent residuals, those residuals' random scatter, or no time series pattern in residuals (Weisent et al., 2010) is always preferred when choosing the best model. In practice, the Q-statistic (also called the Box-Pierce statistic or the Ljung-Box statistic) is used as an objective diagnostic measure of white noise for a time series to compare whether the autocorrelations from residuals and white noise are statistically significantly different. This test statistic is illustrated in Equation 2.11:

$$Q = n(n+2) \sum_{i=1}^{k} \frac{ACF(i)^2}{n-i} \tag{2.11}$$

where k is selected from the lesser of two seasonal cycles, one-fourth of the observations, or 24 when two seasonal cycles are much greater than 24. In most cases, if a model is lacking white noise, this model is deficient and has to be rectified (DeLurgio, 1998).

2.3 Results and Discussions

2.3.1 Employee Turnover Patterns and Outlier Identification

The time-based turnover series was restructured so that it could be analyzed both to observe patterns like trend or seasonality and to understand any inherent time series in the data for further investigation in that direction. As shown in Figure 3.1 (a) the series contains a clear seasonality pattern. The box plot for the turnover series from the ANOVA test confirms this seasonality pattern. In addition, there is a decreasing trend from January to November as shown in Figure 3.1 (b); then the trend line rises in December. The points labeled 2008 in Figure 3.1 (b) are considered outliers since they are beyond the upper whiskers. After all the ARIMA models were tested, three outliers appeared in the SPC chart (as shown in Figure 2.2); the box plot analysis also flagged two of these outliers.

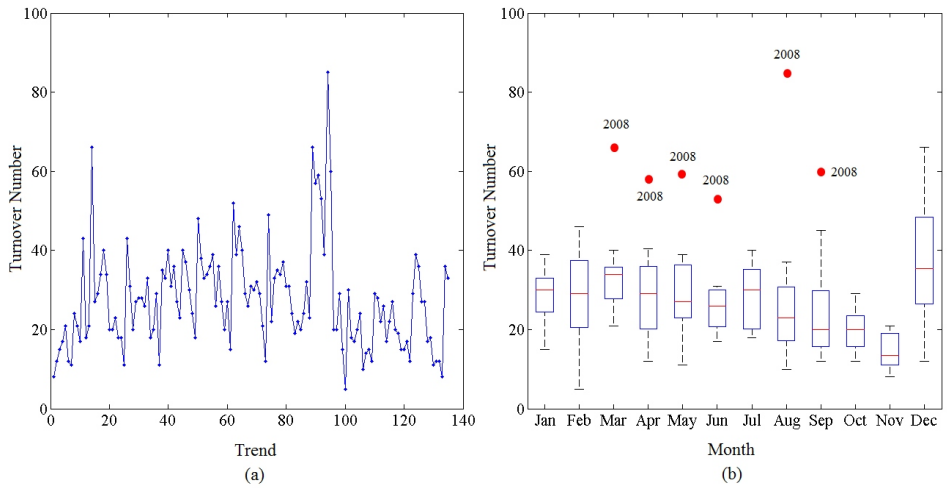


Figure 2.1: (a) Monthly Turnover Series Plotted over Time and (b) Box Plot of Turnover Data

Combining the outliers identified by ANOVA test and SPC, 7 outliers were identified. Those outliers are December 2001, March 2008, April 2008, May 2008, June 2008, August 2008, and September 2008. While December 2001 and March 2008 to July 2008 were treated as a temporary pulse and steps, August 2008 was treated as a pulse with gradual decay in the dynamic regression with intervention methods. In other methods, these outliers were smoothed to soften their impacts, which were univariate regression without intervention,

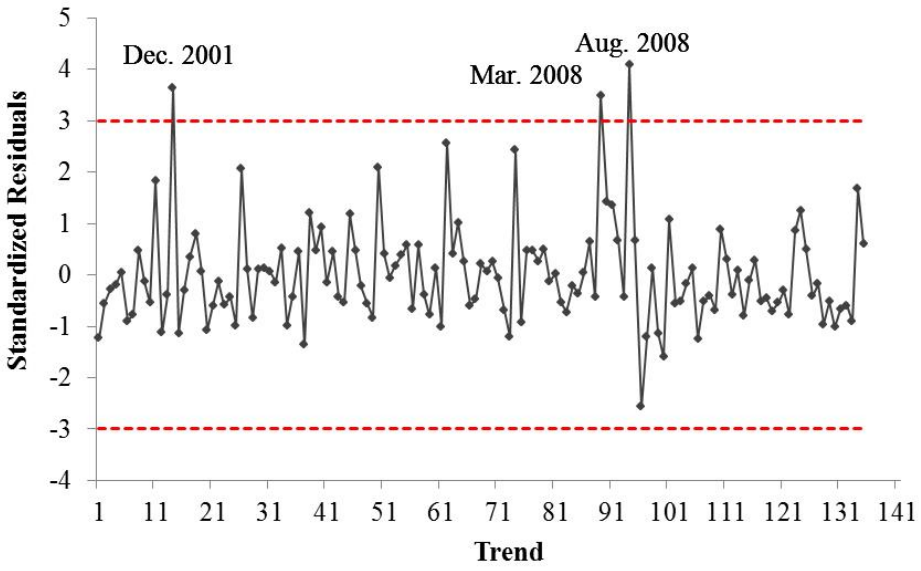


Figure 2.2: SPC Chart for Standardized Residuals

decomposition, exponential smoothing, ARIMA, state space, and vector autoregressive methods. Whenever outliers exist, the cause must be investigated. For instance, December 2001 represents a 9/11 lag impact on job hiring with stronger background checking and increased retiring/hiring. The outliers in 2008 reflect the downsize policy issued in January 2008 with a three months' response-time window to accommodate the organization's voluntary reduction in workforce.

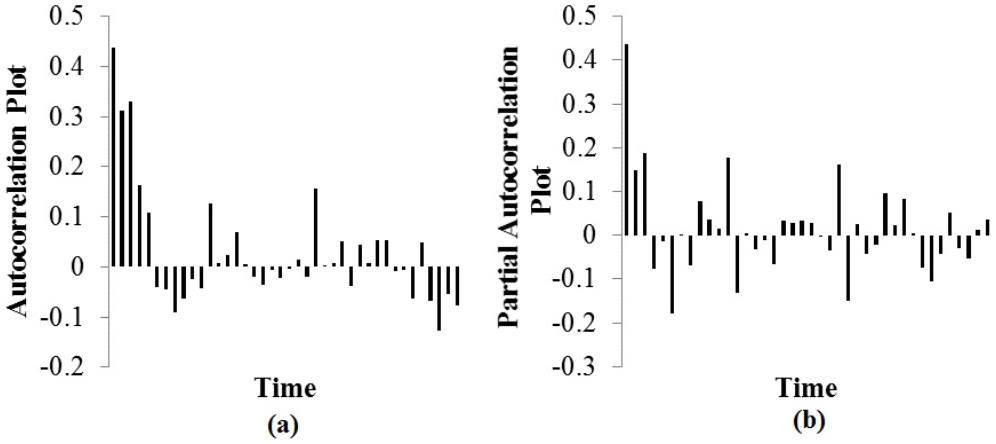


Figure 2.3: (a) Autocorrelation and (b) Partial Autocorrelation Plots

The ACF and PACF plots for the turnover series are shown in Figure 2.3 (a, b). The pattern of unsmoothed data in the ACF and PACF hints at ARIMA(1,0,1) with some type of seasonality, but the seasonal pattern is not obvious.

2.3.2 Cross-correlations

The cross-correlation was applied between the turnover series and the CLI series to identify significant lag correlations. The cross-correlations between first differences for turnover and the CLI series were examined (DeLurgio, 1998), and a pre-whitening process for the two series was used to identify cross-correlation patterns (Box et al., 1970; Bowie and Prothero, 1981). Based on all these calculations, CCFs from Lag 0 to Lag 8 were found to be statistically significant, indicating that the turnover series has a statistically significant correlation with CLI and its 8 lags. When using the cyclical variable for forecasting, the CLI and its 8 lags were incorporated into the dynamic regression, decomposition, and ARIMA model, respectively.

2.3.3 Forecasting Results and Comparisons

Forecasting evaluations for the time series models are provided in the Appendix: Table A.1, Table A.2, Table A.4, and Table A.3. Based on the evaluation statistics, eight univariate models were selected because of an acceptable R^2 value for training and holdout data as well as their residual statistics that are optimum among the other models' statistics (as shown in Table 2.2). On average, the holdout R^2 value of these eight models is 0.51 (ranging from 0.40 to 0.59).

2.3.3.1 Univariate Methods (without External Variables)

The regression model with additive trend and seasonality has the highest holdout R^2 (0.57) among the univariate models without external variables, indicating the model's ability to explain 57% of the holdout sample's total variation. This model is statistically significant ($P < 0.05$) for the model and parameters. However, the residuals are not normally distributed, and the model does not pass the white noise test.

Table 2.2: Statistics for Selected Time Series Univariate Models

Method	#	Model	Pred R^2[1]	Holdout R^2	MAPE	Normality[2]	WN[3]
Univariate without external variable	U1	Regression with additive trend and seasonality	0.51	0.57	26.15	No	No
	U2	Regression with additive trend, seasonality and intervention4	0.72	0.52	22.84	Yes	No
	U3	Decomposition	0.65	0.54	17.97	Yes	Yes
	U4	WES with additive trend and seasonality	0.52	0.52	20.65	Yes	Yes
	U5	ARIMA(1,0,1)(0,1,1)	0.47	0.40	22.89	Yes	Yes
Univariate with external variable	V1	Dynamic regression using lag7 CLI as predictor[4]	0.77	0.59	19.91	Yes	Yes
	V2	Decomposition using lag1 CLI as cycle	0.65	0.55	17.97	Yes	Yes
	V3	ARIMA combining with lag1 CLI as cycle	0.37	0.41	22.73	Yes	Yes

[1] Pred. R^2 is a predicted R^2 value for training data.
[2] Normality is residuals' normality test.
[3] WN is white noise test
[4] The data is unsmoothed (or outliers are unadjusted) so as to take advantage of time series models that can accommodate interventions.

The regression model with additive trend, seasonality, and interventions (pulse and step) performs well with a training R^2 of 0.72 and a holdout R^2 of 0.52, indicating the model's ability to explain 72% of the training sample's total variation and 52% of the holdout sample's total variation. This model is statistically significant ($P < 0.05$) for the model and parameters. The model has normally distributed residuals, but does not have white noises. However, this regression model captures the spike in December 2001 and sharp fluctuations from March 2008 to August 2008 as shown in Figure 2.4.

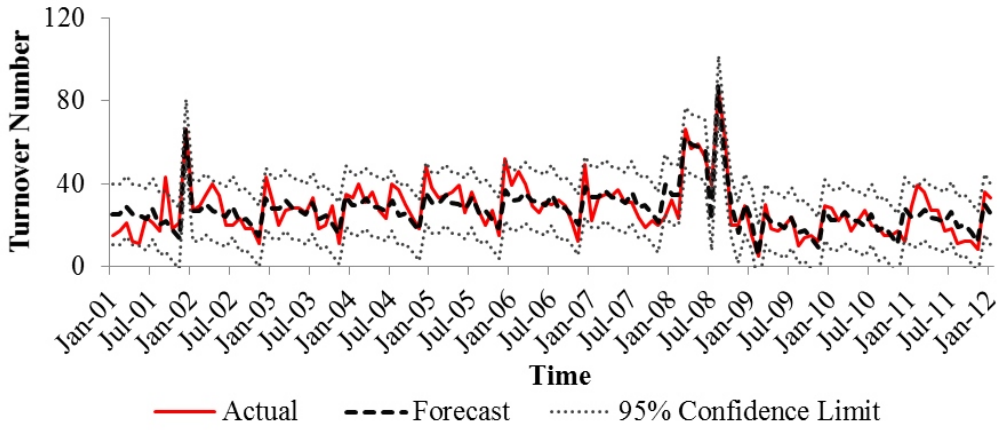


Figure 2.4: Regression with Interventions Actual vs. Forecast Turnover Number Plot

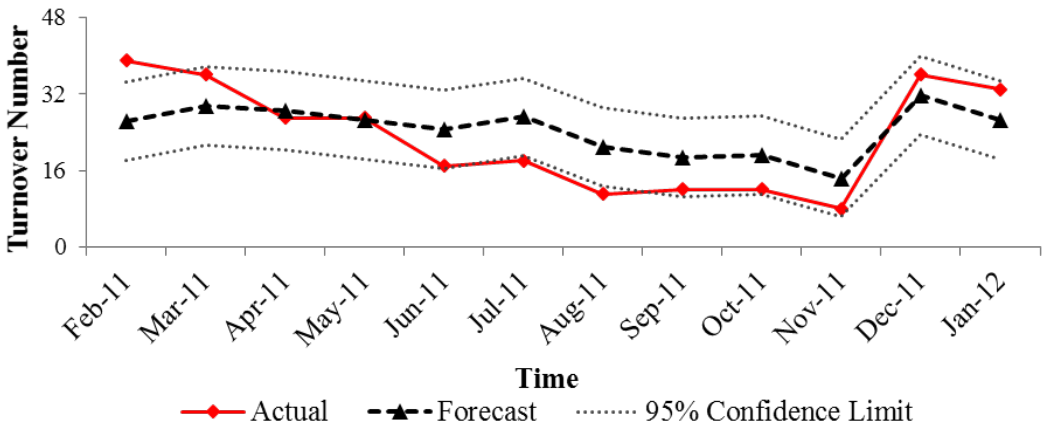


Figure 2.5: Decomposition Actual vs. Forecast Turnover Number with Prediction Intervals for Holdout Dataset

The decomposition model was considered as the best univariate model without external variables because this model has a reasonably high training R^2 value (0.65), a good holdout R^2 value (0.54), and a low MAPE (17.97). This model's residuals are normally distributed and have a white noise pattern. Figure 2.5 plots the predicted turnover based on the decomposition model and the actual turnover number for the holdout dataset. This plot validates the decomposition model's holdout performance as it mimics the changes in trend and the turnover's seasonality. Furthermore, the prediction is close to the actual turnover numbers. However, this model seems to over-forecast for the six months June through November.

2.3.3.2 Univariate Methods (with External Variables)

According to the cross-correlation analysis result, CLI and its 8 lags were applied in dynamic regression, decomposition model, and ARIMA(1,0,1)(0,1,1), respectively as an external variable to forecast turnover number. The dynamic regression model with additive trend, seasonality, interventions (pulse and step), and lag7 of CLI is the best model because it has highest predicted and holdout R^2 value (0.77 and 0.59), normal residuals, and white noise. The dynamic regression is globally statistically significant and individually significant for the parameters ($P < 0.05$). Figure 2.6 shows the predicted and actual turnover number plots for holdout dataset from the dynamic regression model. Although there was over

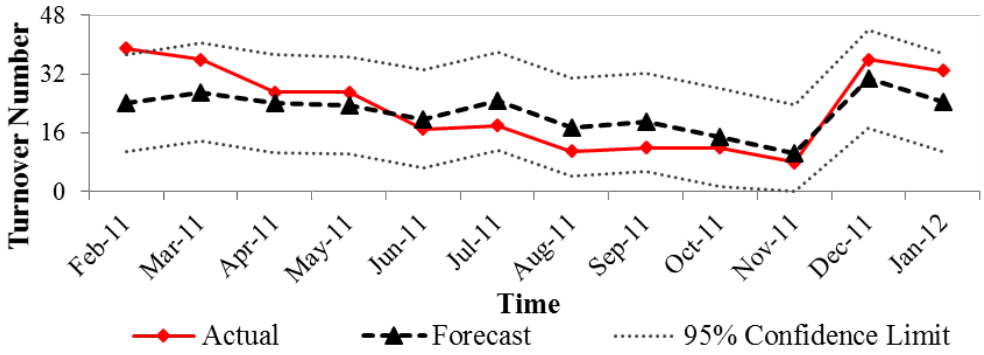


Figure 2.6: Dynamic Regression with Lag7 CLI Actual vs. Forecast Turnover Number for Holdout Dataset

forecasting from July to September as well, the differences between forecasting values and actual values were much smaller. Compared with top-rated univariate methods without external variables, the dynamic regression model's performance was much improved after using CLI as an outside cyclical variable.

2.3.3.3 Nonlinear and Multivariate Methods (with External Variables)

State space models with various combinations of errors, trend, seasonality, or an exogenous variable (CLI here) and VAR models of bivariate time series (turnover and CLI) were employed to forecast turnover number (as shown in Table 2.3). The exponential smoothing state space model with multiplicative error, no trend, and multiplicative seasonality had the highest holdout R^2 value (0.43) among all state space models, which the R `forecast` package automatically selected from 27 exponential smoothing state space models. The residuals have white noise and are normally distributed. However, the training and holdout R^2 values (0.53 and 0.43, respectively) are relatively lower than the univariate regression models .

Table 2.3 shows that the VAR models perform better than state space models. The VAR (4, constant, trend, and seasonality) was considered as the best among all these nonlinear and multivariate model with the highest training and holdout R^2 values (0.62 and 0.64, respectively). The residuals are normally distributed and have white noise. However, variables in this model (lag terms of turnover and CLI) are not statistically significant,

Table 2.3: Statistics for Selected Nonlinear and Multivariate Models

Method	#	Model	Pred. R^2	Holdout R^2	MAPE	Normality	WN
State Space	S1	Trend, Seasonality	0.27	0.39	24.80	No	Yes
	S2	Trend, Slope, Seasonality	0.23	0.32	28.58	Yes	Yes
	S3	Trend, Slope, Seasonality, CLI as regressor	0.13	0.32	28.89	Yes	Yes
	S4	Exponential Smoothing (M,N,M) [1]	0.51	0.43	21.14	Yes	Yes
	S5	Structural TS (Level, Slope)	0.46	-0.16	22.82	No	No
	S6	Structural TS (Level, Slope, Seasonality)	0.51	0.04	21.47	Yes	Yes
Vector Autoregressive [2]	VA1	VAR(5, Constant, Seasonality)	0.62	0.49	18.20	No	Yes
	VA2	VAR(5, Trend, Seasonality)	0.63	0.52	18.37	Yes	Yes
	VA3	VAR(1, Constant, Trend, Seasonality)	0.52	0.47	21.08	No	Yes
	VA4	VAR(2, Constant, Trend, Seasonality)	0.59	0.44	19.63	Yes	Yes
	VA5	VAR(3, Constant, Trend, Seasonality)	0.61	0.51	19.06	Yes	Yes
	VA6	VAR(4, Constant, Trend, Seasonality)	0.62	0.64	18.44	Yes	Yes
	VA7	VAR(5, Constant, Trend, Seasonality)	0.65	0.61	17.98	Yes	Yes

[1] M, N, M is multiplicative errors, no trend, multiplicative seasonality, respectively.
[2] VAR(p) is pth order of lag term.

indicating the model is over fitted. Only VAR (1, constant, trend, and seasonality) has significant lag 1 term of both turnover and CLI. This model also has higher training and holdout R^2 values (0.57 and 0.47, respectively) and white noise. However, its residuals are not normally distributed.

The bivariate time series (turnover and CLI) does not have a multivariate moving average pattern. Thus, it is not statistically significant. Therefore, the vector moving average (VMA) and vector autoregressive moving average (VARMA) methods were not used to forecast. The volatility models (Garch (1,1) and stochastic volatility model) and nonlinear models (nonlinear autoregressive model and nonlinear threshold autoregressive model) were also considered. However, all these models have lower training and holdout R^2 values and higher MAPE values. Therefore, these models' statistics are not provided in Table 2.3 due to their poor performance.

Even more non-linear and multivariate models were considered, but the intent of this research was not to search for the best model but to find a simplistic one that human resource management (HRM) could use. Noteworthy is that a combination model might have been better than the dynamic regression model, but this study was trying to keep the model as simple as possible for HRM.

2.4 Conclusion

In this study, various time series forecasting models for predicting employee turnover were tested, and optimal models for turnover forecasts were identified. As a result of the external variable, the model in this study actually performed better than those accessed in the literature review. Although VAR (4, constant, trend, and seasonality) has the highest holdout R^2, normally distributed residuals and white noise, the dynamic regression model is considered the best forecasting model. Univariate methods were selected for several reasons. Compared to univariate models, multivariate models help in generating a more accurate model fit in most cases. However, univariate models are preferred as they negate several drawbacks of multivariate models. For example, univariate models have less parameter uncertainty and less chance for outliers and errors because of their design simplicity. In most cases, univariate models are also easier to develop, interpret and get concrete conclusions. In contrast, multivariate models are more susceptible to misspecification because of their complexity. Furthermore, in univariate models' explanatory variables have to be determined accurately before forecasting the dependent variable. Errors in forecasting the explanatory variable for a multivariate model may significantly affect the dependent variable's forecasting accuracy when compared with that of an equivalent univariate model (Chatfield, 2000). Apart from the benefits univariate models' benefits, the dynamic regression model has several additional advantages. For example, an ARIMA error term, which has an autocorrelation pattern, can be included in the model. The dynamic regression model can also handle lagged regressors and various types of seasonalities. In addition, the dynamic regression model can effectively handle interventions or change points (such as holidays, promotions, and new policies) since they are often common in the time series data.

Thus, the dynamic regression model could be used to forecast turnover for most organizations of any size. However, in implementing dynamic regression modeling, at least five years of monthly employee turnover data is preferred to make an accurate forecast. If the dataset's horizon is less than five years, a special decomposition model (Ittig, 1997) could be considered as a substitute. Although the dynamic regression model's forecasting horizon is relatively short, this is not a big issue for human resource departments since most are only interested in a short-term, such as three months, forecasting. Therefore, if an organization's human resource (HR) department is unfamiliar with forecasting techniques, the dynamic regression model could be a good option for a preliminary turnover forecast once a CLI is identified.

Noteworthy is that an external variable, such as CLI in this study, helps in forecasting turnover since it anticipates cyclical turning points. Incorporating such an external variable in a model is very helpful in getting a good forecast when the HR department has a small and unreliable data set. Incorporating external variables, such as CLI, may also help the entire forecasting process. If an external variable such as CLI is unavailable , a decomposition model could be considered as the first choice rather than a dynamic regression model. In practice, some software, such as NCSS or MINITAB, has an embedded decomposition macro, which HR departments could easily run because they know how to estimate cyclical variation.

2.4.1 Practical Implications

According to this study's findings, employee turnover forecast, could be handled easily. HR departments could use a univariate linear regression model for the preliminary forecast, whose accuracy is acceptable. However, regression cannot handle some types of interventions, such as pulse or steps with exponential decay or growth. Dynamic regression could be used as an alternative for forecasting in such cases.

In this study, statistical analysis packages SAS and NCSS were used for forecasting. However, when HR departments are unwilling to devote extra funding to software purchases, Microsoft Excel could be a good alternative for the time series forecast because open-source time series forecasting packages have been designed to run in the Excel environment. Forecasting models (such as naive, moving average, exponential smoothing, decomposition,

regression, and ARIMA) have been included in these packages (Warren, 2008). Another option is to use R for the dynamic regression (Hyndman, 2014).

2.4.2 Limitations and Future Research

This study was limited to forecasting turnovers in an organization. Time series forecasting models could also be used to forecast turnovers in retirement or voluntary resignations. Although this study used CLI as an external factor and the forecasts' accuracy significantly improved, the external factors affecting turnover were far beyond CLI's scope because local and cyclical economic fluctuations also strongly influence employees' propensity to quit (Abelson and Baysinger, 1984). Also, an individual level turnover forecasting model is crucial for the future study. It assists HR manager to identify the high turnover area in the organization and to target the employees with high turnover probabilities using survival analysis techniques.

Chapter 3

Employee Turnover Forecasting and Analysis Using Survival Analysis

3.1 Introduction

Employee turnover is a topic that has drawn the attention of management researchers and practitioners for decades because it is both costly and disruptive to the functioning of most organizations (Staw, 1980; Mueller and Price, 1989; Kacmar et al., 2006). Human resource analytics systems' key goals include identifying not only factors leading to employee satisfaction and productivity but also turnover's causes and timing (IBM, 2013). For many mature firms with large workforces, an important piece of this puzzle is developing predictive models for both retirement and quitting. While commercial tools may exist in this space, very little discussion of applied predictive models has appeared in academic literature. From an operational perspective, the ability to accurately predict turnover across a range of organizations and job types is a highly beneficial to front-line management of these organizations and to their financial, human resources, and actuarial concerns of the company and its supporting partners. The ability to forecast turnover becomes even more valuable in specialized industries and government agencies with long hiring lead times. From a research perspective, a predictive retirement model is a platform that can allow investigators to evaluate both external economic and demographic factors as well as internal policies influencing retirement decisions.

This study focuses on the behavior of individuals between 2000 and 2012 employed by a large industrial organization located at a single site that provided employees with a defined benefit retirement plan. The study has four objectives: 1) Develop a probabilistic model of the employee lifetime as a function of basic demographic, employment, and external factors. 2) Evaluate the model's aggregate predictive accuracy for one-and two-year time frames as a tool to facilitate planning. 3) Determine the internal and external economic variables' impact on retirement. 4) Quantify an early retirement incentive's impact on retirement behavior. Because of the sampling approach used to collect the data, an integral part of the study was ensuring that the modelling strategy was robust to biases introduced by left truncation and right censoring.

To analyze the data, the Cox proportional hazards models (PH) was used. The strength of the Cox model is the semiparametric form incorporating a non-parameteric baseline estimate and a parametric term that determines the other factors' relative impact. As discussed in Section 3.4.2, this model is not sensitive to truncation bias, can be viewed as a modeling approach for non-homogeneous Poisson processes and can include the covariates' effects that change over time.

3.1.1 Motivation for the Study

The current work bridges the gap between factors involved in individual retirement decision-making and workforce management's importance from the perspective of human resources. Here, the study's focus is on predicting turnover for strategic planning in large organizations, such as government agencies, large corporations, and large academic institutions in order to determine changes in workforce size and to plan for the eventual loss of critical skills. This effort is particularly applicable to corporations that have an older workforce with many employees close to retirement and most with defined benefit retirement plans. Human resources departments, benefits managers, and actuaries can also use these results to better manage investment resources and plan for the near term.

In many industries with older worker populations, retirement is a major source of HR disruption, causing delays and other problems in processing the work flow. Replacing retired workers can also be a major expense both in HR staff time and recruiting costs. Effective

predictive models can help managers identify potential individuals from which department in an organization likely to leave, thus giving the organization more lead time in planning for and recruiting replacements. For example, in large organizations that use skilled workers in both white-collar and blue-collar jobs, accurately forecasting openings in the next 6 to 12 months can be extremely beneficial in maintaining continuity of operations. Similarly, other firms with younger employee populations may be more affected by employees' quitting and would benefit from predictions.

While aggregate forecast models of turnover and, more specifically, retirement exist (as discussed in Chapter 2, such models have limited ability to estimate at division or job-category levels. Such models also do not consider the population's demographics, such as age, years of service, pension type, and potentially numerous other factors that can influence the probability of retirement. By modeling the distribution of time until retirement at the individual level, much more relevant information can be included, such as age, type of retirement plan, job classification, organizational division, years of service, pension benefits' details, and individual survey responses if they exist, as well as external social and economic trends that are geographically relevant. This model could not only provide accurate predictions but also give managers and researchers feedback on how different factors and incentives may influence retirement and other HR decisions, such as early retirement incentives.

3.2 Literature Review

Employee turnover is a general term referring to the loss of employees resulting from a wide range of causes, such as retirement, death, quitting, termination, promotion, and reassignment. Each of these turnover modes has different foundational causes and may be more or less prevalent during different points in one's career.

A considerable amount of work has been done identifying underlying causes of turnover both from the perspective of retirement and of voluntary quitting. Rainlall (2004) reviewed theories and summarized crucial factors affecting employee retention and turnover both personally and organizationally. The factors investigated involved employee needs in terms

of individual, family, and cultural values, work environment, responsibilities, supervision, fairness and equity, effort, and employees' development. Mal-functional leadership is key factor leading to turnover when managers do not provide appropriate training, facilities and leadership. From the retirement's point of view Wang and Shultz (2010) summarized key theoretical and empirical research studies on causes of retirement between 1986 and 2010 and identified inconsistent findings drawing on research from a wide variety of social science fields. Most of the work cited in that review focuses on the process of retirement and factors driving the retirement decision from the individual perspective. Most relevant to the current work is research in retirement decision making and human resource management.

This study focuses on forecasting, which may be accomplished at the aggregate level or may be broken down by organizational factors or by the mode of loss. For example, using a similar data source considered here, a time series approach to predict future aggregate turnover based on losses in previous years was discussed in Chapter 2. Such an approach's weakness is that it does not use the employee population's known characteristics, such as age, skill set, performance evaluations, salary, years of service, and numerous other factors to attempt to predict turnover. Based on the theoretical and empirical findings described in earlier studies (Rainlall, 2004; Wang and Shultz, 2010), using the factors identified as critical in the retirement and quitting decision-making process, it was expected to improve the ability to predict at both the individual and aggregate levels.

Regression models for lifetime data offer the potential to use internal human-resource data when making predictions. In the organizational and business settings, both academic and professional researchers have used these methods to solve practical problems in the industry. While actuarial scientists have used these methods since their inception to create models in risk and insurance (Brockett et al., 2008), researchers in finance have more recently explored using these models to reflect lifetimes of banks (Lane et al., 1986), as well as time until default of financial instruments such as fixed-income securities (LeClere, 2005).

More relevant to the current work is researchers' application of survival analysis and lifetime data methods to customer relationship management. For example, Lu (2002) applied survival analysis techniques to predict customer churn for cellular phone services. Their study provided a tool for telecommunications companies to design retention plans for reducing

customer churn. Also, Braun and Schweidel (2011) used a hierarchical competing risks analysis to model when and why customers terminate their service by employing the data from a land-based telecommunication services provider.

Until now, however, little has been done in the area of human resources. While Berger and Chen (1993) considered statistical modelling of tenured faculty's retirement within a university setting, they applying the Bayesian statistical approach to modelling retirement outcomes. More recently, major analytics consulting firms, such as IBM and PWC, have begun offering human resource analytics software and services (IBM, 2013; PWC, 2015). Within this area, some consultants have proposed basic survival models for employee churn (Briggs, 2014); but few details are available, and the complexities of real-world situations tend to be avoided.

Survival analysis has been most frequently used to examine lifetime data-generated engineering (reliability) (Lawless, 2011; Meeker and Escobar, 2014; Holt, 2011); medicine and epidemiology (Kalbfleisch and Prentice, 2011); and less frequently, social sciences (event studies) (Allison, 2010; Long and Freese, 2006). The first two fields have motivated most of the theoretical development in these fields. In the engineering reliability area, Carrión et al. (2010) estimates the time to failure of the pipes in a water supply network's dataset under left-truncation and right-censoring by using the extended Nelson estimator (Pan and Chappell, 1998).

These methods have been applied to thousands of epidemiological studies, retrospective biomedical studies, and clinical trials over the past 40 years. For example, Claus et al. (1991) investigated the familial risk of breast cancer in a large population-based, case-control study using recurrent lifetime analysis. Those researches found that the risks of breast cancer are a function of women's age. Moeschberger and Klein (2003) provided a thorough book-length overview of the methods and included specific case studies focusing on medical applications.

3.3 Data Preparation

Provided by a large multipurpose research organization in the U.S., the analyzed dataset consisted of 4316 active and 3782 former full-time employees. This population of employees

was followed across a 12-year window from November 2000 to December 2012. Records of employees who retired or left before November 2000 or who were hired after December 2012 was truncated from the dataset. In addition, 4316 current employees had no termination date. The sampling approach taken, capturing only employees active in a fixed window, created two forms of bias-right censored and left truncation-in the sample that must be accounted for. Subjects were right censored if their endpoint (retirement and quitting in this case) was unknown at the time of the study since they were still actively employed. Right-censored observations provide information and should not be dropped but must be analyzed differently than complete observations. Left truncation results from a failure to include cases that failed before the study window's beginning. A biased sample resulted because only those cases surviving long enough were represented in a sample. Both of these potential biases are considered in the discussion of models (see Section 3.4.1). Several static employee attributes are provided in the list below:

- Payroll (PR): hourly, weekly, or monthly payroll
- Gender (GENDER): male, female
- Division (DIV): used to distinguish the departments, among ten departments. In this study, each division's definition was not static throughout the entire observation period. Over the course of time, divisions could be renamed, reduced, or dismissed in reorganizations. Furthermore, no employees transfer between divisions was recorded. Therefore, for prediction purposes, the division indicates the organization level that an employee was associated with at the time of the final observation.
- Occupational Code (OC): a standardized code used to describe the job category within the organization for reporting purposes. These codes include Crafts(C), Engineers (E), General Administrative (G), Laborers (L), General Managers (M), Administrative (P), Operators (O), Scientists (S), and Technicians (T). In this study, occupational codes are highly correlated with payroll category: managers, engineers, administrative, and scientists are on monthly payroll; general administrative employees and technicians are on weekly payroll; and other categories are paid on an hourly basis.

- Age at Hire (AGEH): either the age of an employee when most recently hired or that employee's age in November, 2000.
- Age at Credit (AGEC): age of an employee at most recent time that employee receives credit for pension
- Years of Current Service (YCS): years of service accounting for pension credit
- Years of Service at Hire (YCSH): years of service accounting for pension credit at the employee's most recent hiring
- Termination Date (TD): the date when an employee left the organization
- Termination Type: an employee's reason for leaving the organization, such as retirement (RE) or voluntary quitting (VQ).
- Points: the sum of the employee's age and YCS. (YCS can be larger than 0 at hire if an employee has credit from earlier employment in the same organization.)

These twelve indices were operationalized for testing using a 12-month lag of their one-year averages. This approach ensured that the variable could be useful in forecasting since the one-year lag was known at the time of forecasting. The economic indices were originally reported daily or monthly. The yearly average was computed as the index's average value over the previous 12 months.

3.4 Model Development and Evaluation

This study aimed to develop accurate predictive models of retirement and quitting behavior. The models were used to address several key questions:

1) How accurately can retirement (quitting) be predicted?

2) What factors indicate an individual is more likely to retire (quit)?

3) Which external economic factors are most predictive of retirement (quitting)?

4) What is the magnitude of the impact of an Early Retirement Incentive Programs (ERIP)?

5) How do the tenure and age impact retirement?

6) How many employees, by occupational category and division, will retire (quit) next year?

Survival or lifetime data analysis is used to study the time distribution required for a subject from a population to experience an event such as mechanical failure, death, or recovery. Survival regression models relate lifetime distribution's factors, such as the hazard function, to a linear function of explanatory variables. Statistical survival models are often separated into two categories: parametric survival models and semi-parametric proportional hazards (PH) models or Cox models. In this study, the Cox PH model was employed to build predictive models of retirement and quitting, to estimate an employee's baseline hazard of retirement or quitting, and to identify significant factors that might impact turnover. The parametric models were inappropriate for this study for several reasons. First, it was unlikely that hazards for events like retirement would match common parametric distributions, such as Weibull or log-normal, since the risk should remain close to zero until the usual range of retirement, when it spikes and then drops quickly again. Furthermore, as mentioned in the introduction, the sampling scheme used in this data involved several biases, which could most easily be adjusted to use the Cox PH model. The Cox model's third advantage is the ability to incorporate time-dependent covariates. In the case of the current model, these covariates were required both incorporating the impact of a 2008 early retirement incentive program (ERIP) into the organization and examining the effects of two other variables related to pension benefits. Financial indices that vary with time were also captured using this methodology. A special version of the model known as the competing risks analysis was applied for modeling a population who could experience two types of events (in this case, employee retirement and voluntary quitting). In addition to the model fitting, a simulation study was performed to examine the impact of data bias on the Cox proportional hazard model's forecasting capability.

3.4.1 Missing Data Biases: Right Censoring and Left Truncation

Right censoring and left truncation are commonly observed forms of missing data in survival analysis data set. In the current study, the study window was from November 2000 to

December 2012 as shown in Figure 3.1. The complete records of active employees during

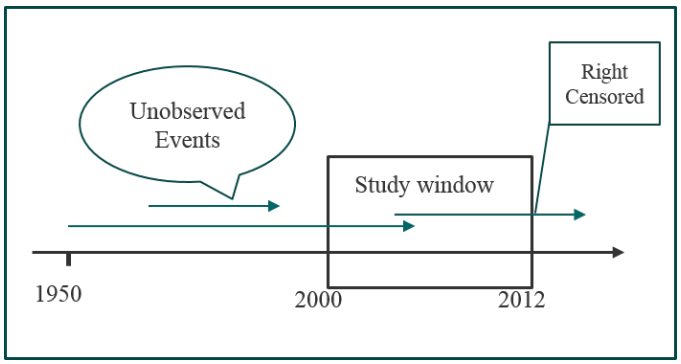


Figure 3.1: Right Censoring and Left Truncation

this period were included in the dataset whether or not those employees' tenure began or ended outside the study window. Conversely, those employees whose tenure ended before the study window or whose start date occurred after the study window ended were not included in the study.

Let T be the time at which an individual experiences the event of interest and let C denote the final time the individual is observed. An observation is called *right censored* if $T > C$, indicating that actual event time for the individual is not recorded but is only known to be greater than C. Thus, employees active at the end of the observation window are right censored. Right censored observations contained information, although incomplete, about a subject's lifetime and require special treatment in order to draw a proper inference,

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq c_i \text{ (uncensored)}, \\ 0 & \text{if } t_i > c_i \text{ (censored)}, \end{cases}$$

where i denotes the i^{th} observation, and the event's failure time for i^{th} observation is the minimum time between t_i and c_i, i.e., $min(t_i, c_i)$, that is when $c_i < t_i$, c_i is taken as the i^{th} observation's end time in order to do the next analysis.

Left truncation is another interesting artifact of the window sampling scheme. Let T again denote the time that the event of interest occurs, and let X denote the time an individual enters the study. Only the individuals with $T \geq X$ are observed in the study window. Those individuals with $T \leq X$ are referred to as left truncated because they could

not be included in this study based on the sampling window as shown in Figure 3.1. Left truncation exaggerates the number of longer-life individuals, thus leading to a biased sample as shown in Figure 3.1. The longest arrow represents a life span of an employee who was hired in 1950 and retired in 2006. While this employee and any in his cohort who are still active are in the sample, others that began in 1950 but retired in 1998, for example, are not. Hence, in the sampling-window approach the longer someone continues working, the more likely they are to appear in the dataset. Therefore, an overabundance of longer living individuals is seen in the data. Left truncation's presence and the associated bias in the data must be considered to accurately estimate survival (Carrión et al., 2010).

3.4.2 Cox Proportional Hazards Regression Model

The Cox proportional hazards (PH) regression model is the most widely used method for modelling lifetime data. Introduced in Cox's (1972) seminal paper (one of the most cited papers in history) (Cox, 1975), the Cox PH model is the canonical example of the semi-parametric family of models, specifying a parametric form for the covariates' effect on an unspecified baseline hazard rate, which is estimated non-parametrically. The form of hazard model formula is shown in Equation 3.1:

$$h(t, x) = h_0(t)e^{(\sum_{i=1}^{k} \beta_i x_i)} \tag{3.1}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ are characteristics of individual i, $h_0(t)$ is the baseline hazard, and β is a vector of regression coefficients. The model provides an estimator of the hazard at time t for an individual with a given set of explanatory variables denoted by x_i. In the standard Cox model, the linear combination $\sum_{i=1}^{k} \beta_i x_i$, is not a function of time t, and is called time-independent. If $x_i(t)$ is a function of time the model is called the extended Cox PH model which is discussed in Section 3.4.3. A key assumption for the model is the proportional hazards assumption, which assumes that explanatory factors have a strictly multiplicative impact on the hazard function so that different groups maintain a constant hazard ratio at all times. However, the Cox PH regression can be extended to handle non proportional hazards using time-dependent variables or stratification; see Klein and Moeschberger (2003).

The Cox PH regression is "robust" and popular, because the baseline hazard function $h_0(t)$ is an unspecified function and its estimation can closely approximate the correct parametric model (Kleinbaum, 1998). Taking both sides of the equation's logarithm, the Cox PH model is rewritten in Equation 3.2:

$$\log h(t, x) = \alpha(t) + \sum_{i=1}^{k} \beta_i x_i \tag{3.2}$$

where $\alpha(t) = \log h_0(t)$. If $\alpha(t) = \alpha$ (i.e. constant), then the model reduces to the exponential distribution. As noted earlier, the general Cox PH model puts no restrictions on $\alpha(t)$. The partial likelihood method is used to estimate the model parameters (Allison, 2010).

3.4.3 Time Dependent Variable and Counting Process

Some explanatory variable values changed over the course of the study. The extended Cox PH regression is a modification of the model, incorporating both unchanging time-independent variables as well as variables that change with time or time-dependent variables,

$$h(t, x) = h_0(t) e^{(\sum_{i=1}^{k_1} \beta_i x_i + \sum_{j=1}^{k_2} \gamma_j x_j(t))} \tag{3.3}$$

where $x = (x_1, x_2, \ldots, x_{k_1}, x_1(t), x_2(t), \ldots, x_{k_2}(t))$, $h_0(t)$ is the baseline hazard occurring when $x = 0$, β and γ are the coefficients of x. To fit this model, modifying the partial likelihood is required, and the data set is often presented in a format called the *counting process* format in order to facilitate this calculation. The current study considers three internal time dependent variables, which are functions of the individual's characteristics. These variables are described below.

- Early Retirement Incentive Program (ERIP): a specific time during the study window, part of the 2008 calendar year, when the organization offered an early retirement incentive program; see Clark (2002). This program was time varying in the sense that it occurred at a different age for each individual in the study population and was set at 0 during the period when no program existed and at 1 during the period when the program did exist.

- Points 85 (P85): an indicator that an employee amassed 85 service points, the sum of the employee's years of service and age, thus qualifying that employee for full retirement benefits

- Age at 65 (A65): an indicator that an employee qualified for retirement by exceeding the age 65 threshold.

P85 and A65 are time-varying variables capturing important changes in an individual's hazard level throughout the study. The counting process format allows software packages to handle time-dependent variables by creating multiple intervals for each employee. Each interval is defined so that the time-varying variables are constant within the interval. For example, for an individual who remained active until the end of the study, achieved 85 points in December 2003, exceeded age 65 in December 2007, and received a retirement incentive in 2008, five records were included for the individual: November 2000-November 2003; December 2003-November 2007; December 2007; January 2008-December 2008; and January 2009-December 2012, the end of the study.

Financial indices represent another form of time-dependent variables. These indices are referred to as external variables because they depend upon factors external to the employee. In models considering financial indices, monthly observations of these indices are aggregated at the yearly level, leading to a smaller number of time intervals in the counting process format. Economic variables are included at a one-year lag since retirement and quitting decisions are assumed to occur significantly in advance of the actual event and therefore depend on these indicators' older values. Importantly, using lagged quantities allows the model to be adapted to forecast up to 12 months. In the case of data with external indices, the intervals' start points are defined as max (i.e., hired date, January 1st of a certain year) and the end points are defined as min (i.e., terminated date, December 31st of a certain year).

3.4.4 Stratification and Multiple Baselines

An alternative for handling non-proportional hazards is stratification. A stratified model allows each data subgroup as defined by a grouping variable to have its own baseline hazard

while sharing other variables' parameters. If the proportional hazards assumption holds within these subgroups, then this model produces valid common estimates of variable effects using all the observations. Equation 3.4 represents the hazard function for stratum z;

$$h(t, x, z) = h_0^z(t)e^{(\sum_{i=1}^{k} \beta_i x_i)} \tag{3.4}$$

where z represents the grouping variable, and $h^z\sigma_0(t)$ is a baseline hazard based for stratum z and β_i are variables' common effects. Note that the strata variables cannot be the variables in the Cox PH model.

3.4.5 Testing the PH Assumption

Three common approaches are available for testing the proportional hazard assumption's validity. The first approach is investigating the Schoenfeld residuals. A second approach is testing the interaction between time-dependent and time-independent variables in the Cox PH model. The PH assumption is valid if the interaction is not statistically significant ($P > 0.05$). Finally, including separate baseline hazards for each stratum that the analyst defines can also capture variation in the hazard rate's changes. See Allison (2010); Collett (2015) for more details on these tests.

3.4.6 Competing Risks

One of the many nuances observed within this data set is the fact that currently employees can leave employment in several mutually exclusive ways, including quitting voluntarily, being laid off, being dismissed for cause, transferring, retiring, or being unable to continue because of disability or death. A competing risk is an event whose occurrence either precludes the event of interest from occurring or fundamentally alters the probability that the interest will occur (Tableman and Kim, 2003). A competing risks model is a common approach when studying a single mode of leaving, such as retirement, if subjects at risk may also exit through an alternative mode such as quitting. In the current study, when considering retirement as the event of interest and voluntary quitting as a competing risk, all observations were initially

included in the study and outcomes being a quitting event were treated as censored, allowing the observed work period to be used informatively.

3.4.7 Variable Selection and Model Choice

In the current study two equivalent time measurements were considered as response variables for modelling: AGE in years and YCS. Because of including time-varying explanatory variables and the need to estimate the baseline hazard for purposes of forecasting, the data had to be formulated as a counting process, see Section 3.4.3. After some consideration, age was deemed the better option for analysis because the more condensed distribution of values allowed more accurate the baseline estimates.

The model selection was initiated by considering DIV, AGEH, GENDER and other time-independent variables as well as ERIP, P85, and A65. Non-significant variables were removed using the criteria that p-values should be less than .05 and starting with the largest p-value first, (i.e., backwards selection). This removal continued until only statistically significant variables remained. Stratifying the baseline was tested using the occupational code, which did not improve the model. Finally, external time-varying covariates were tested one at a time to capture economic factors, and the impact on model performance was noted.

3.4.8 Baseline Smoothing

To better predict employee retirement and quitting, baseline smoothing methods were applied to generate a smoothed cumulative hazard function. The Cox PH model generates a non-parametric baseline based on the number of events occurring in the reference group. When no events occur in the certain points t, the baseline function for time t is equal to the value at t^*, when the event occurs right before time t (given $t > t^*$). Thus, when the data have a high proportion of censors, the model will underestimate the failure probability. The cumulative hazard function was smoothed to see whether a better prediction could be achieved. The smoothing method smooths the cumulative hazard function using SAS proc transreg. The new baseline replaces the original non-parametric baseline and generate the failure probability for each individual. In using a smoothed baseline, the number of

predicted events is compared with the original baseline's results. Because the voluntary quitting data has a high proportion (90%) of censor, the number of predicted events may be more accurate after using a smoothed baseline. On the other hand, since one-third of the terminated employees are retired, the smoothed baseline methods may not make better predictions than the original one; however, it may build a better estimation of the cumulative function at the retirement baseline's tail.

3.4.9 Model Evaluation and Comparison

To evaluate the models considered in this study, the data were first split into two sets: a training set containing all the observations from 2000-2010, and a testing set containing events that occurred in 2011 and 2012 and that involved the same individuals. The testing (holdout) sample was included to accurately measure how well the model would forecast beyond the observed data. Because the testing data set was not included in the model-fitting process, this out-of-sample evaluation better estimates predictive accuracy.(see Kuhn and Johnson (2013) for further discussion of this approach.)

All the fitted models considered in this study were first evaluated using four statistical criteria: Akaikes Information Criterion (AIC), Schwartzs Bayesian Criterion (SBC), mean absolute percentage error (MAPE), and likelihood-based goodness of fit G^2. The optimal model should minimize the values of AIC, SBC, MAPE, and G^2 when fit to the training data. In this study, the model performance on the holdout dataset was considered more important than that on the training dataset. AIC and SBC assess model fit by balancing a larger likelihood value with a penalty increasing with the number of variables included. Including the penalty term diminishes the potential for over-fitting (Allison, 2010; Hosmer et al., 2013). In this study, these measures were generated automatically using the model-fitting process.

To asses predictive measures such as MAPE and G^2 , the event of interest's probability was predicted (e.g., retirement) for each active individual during each calendar year of the training or testing data set. For each employee, the conditional probability of the event occurring between t_j and t_{j-1} was computed, given that the employee was active at time t_{j-1}. This probability was calculated using the baseline hazard and coefficients from Cox

PH models as shown in Equation 3.5 below

$$\begin{aligned} P\{t_{j-1} < T < t_j | T \geq t_{j-1}\} &= 1 - P\{T > t_j | T \geq t_{j-1}\} \\ &= 1 - \frac{S_k(t_j)}{S_k(t_{j-1})} \\ &= 1 - \frac{S_0(t_j)^{exp(\sum_{i=1}^{p} \beta_i x_i)}}{S_0(t_{j-1})^{exp(\sum_{i=1}^{p} \beta_i x_i)}} \end{aligned} \tag{3.5}$$

where T_k is survival time of the k^{th} individual , t_j is a specific time value, $S_k(t) = S_0(t_j)^{(\sum_{i=1}^{p} \beta_i x_i)}$ is the survival function for the k^{th} individual, $S_0(t)$ is the baseline function that Cox PH model generated, x_i are the individual explanatory variables, and β_i are the respective regression coefficients for the variables.

The yearly predicted number of events is the sum of conditional probabilities given by Equation 3.5 as shown in Equation 3.6,

$$\begin{aligned} \text{E(Turnover } t_j) &= \sum [1 - \frac{S(t_{ij}|x_i)}{S(t_{ij-1}|x_i)}] \\ &= \sum \left[1 - \frac{S_0(t_{ij})^{exp(\sum \beta x_i)}}{S_0(t_{ij-1})^{exp(\sum \beta x_i)}}\right] \end{aligned} \tag{3.6}$$

where t_{ij} is a specific time value for individual i. The variability in estimates of yearly turnover is the sum of two quantities shown in Equation 3.7,

$$\begin{aligned} \text{Var(Turnover } t_j) &= \text{Var}[\sum Y_i] \\ &= \text{Var}[\sum Y_i | P_i] \\ &= \text{E}[\text{Var}(\sum Y_i | P_i)] + \text{Var}[\text{E}(\sum Y_i | P_i)] \end{aligned} \tag{3.7}$$

given, P_i is whether an employee left the organization. Thus, yearly individual turnover variation was modeled as a Bernoulli distribution, $Y_i \sim Bernoulli(P_i)$, thus, $\text{Var}(\sum Y_i | P_1, ...P_n) = \sum P_i(1 - P_i)$. Therefore, $\text{E}[\text{Var}(\sum Y_i | P_i)] = \sum P_i(1 - P_i)$. Variation occurs in the estimating individual turnover probabilities's vector $p \approx (\hat{p_1} \dots \hat{p_n})$ derived from the Cox model. Thus, $\text{Var}[\text{E}(\sum Y_i | P_i)] = \text{Var}[\text{E}(\sum (1 - \frac{S_0(t_{ij})^{exp(\sum \beta x_i)}}{S_0(t_{ij-1})^{exp(\sum \beta x_i)}} | P_i))]$, given, $S(t|x) = exp[-\Lambda(t|x)]$, where $\Lambda(t)$is the cumulative hazard function, and $\hat{S}$ and $\hat{\beta}$ represent random

variables that are data functions.

$$\begin{aligned}
\text{Var}[\text{E}(\sum Y_i|P_i)] &= \text{Var}\big[\text{E}\big[\sum\big[\frac{e^{-\Lambda(t_{ij}|x_i)}}{e^{-\Lambda(t_{ij-1}|x_i)}}|P_i\big]\big]\big] \\
&= \text{Var}\big[\text{E}\big[\sum\big[e^{(\Lambda(t_{ij-1}|x_i)-\Lambda(t_{ij}|x_i))}|P_i\big]\big]\big] \\
&= \text{Var}\big[\text{E}\big[\sum\big[e^{h(t_{ij}|x_i)}|P_i\big]\big]\big] \\
&= \text{Var}\big[\text{E}\big[\sum\big[e^{h_0(t_{ij})exp(\sum\beta x_i)}|P_i\big]\big]\big]
\end{aligned}$$

In this case, computing $\text{Var}[\text{E}(\sum Y_i|P_i)]$ was difficult because of the covariates' repetitive nature across observations. Bootstrap and simulation methods were employed to compute model estimates' variance. The bootstrap resampling methods for censored survival data was used to compute the model estimates' standard error using R `boot` package (Davison and Hinkley, 1997). Ten thousand datasets were generated using a resampling method involving a high performance computing platform. The data was fitted using the Cox model, and the expected failure number during each year was then computed based on the model. The standard error was computed based on the ten thousand replications; and the 95% prediction confidence intervals were computed assuming that sum statistic is normally distributed, which follows from the Central Limit Theorem.

MAPE is a common measure for computing predictions' accuracy from a forecast model and is often used to compare models since it measures relative performance (Chu, 1998). MAPE is calculated as the average percent deviation of a forecast from the actual observation as shown in Equation 2.10. The implementation of MAPE for predictions of retirement and quitting used the yearly actual and predicted numbers of events (y_t and $\hat{y}_t$ respectively). The predicted yearly retirement number, $\hat{y}_t$, was the expected retirement count for a particular year and was computed as the sum of conditional probabilities expressed in Equation 3.5 over currently active employees. This number follows from the fact that each employee's probability of retiring could be viewed as an independent Bernoulli random variable.

Although less common in forecasting, G^2 is another useful criterion for evaluating a model's prediction of dichotomous events (Simonoff, 2013). The calculation takes the form,

$$G^2 = 2\sum_t [y_t \log(\frac{\bar{p}_t}{\hat{p}_t}) + (n_t - y_t)\log(\frac{1-\bar{p}_t}{1-\hat{p}_t})] \tag{3.8}$$

where y_t is the number of events in year t, n_t is the workforce number in year t, $\bar{p}_t = y_t/n_t$ is the observed proportion of events, and $\hat{p}_t = \hat{y}_t/n_t$ is the model's predicted number of events. Small values of G^2 indicate close agreement of observed and predicted numbers of events.

3.5 Proportional Hazards Models' Simulation Studies

The goal of this data analysis was to create a predictive model for retirement and other types of turnover based on an of employees database. Section 3.4.1 pointed out two sources of bias present in the data: left truncation and right censoring. Because the Cox PH model is estimated using a partial likelihood that depends only on the cases at risk at the specific failure times, estimates of regression coefficients should remain unbiased and efficient in the presence of left truncation and right censoring (Harrell, 2013). What is less clear is the impact of the bias on model predictions because they depend on the baseline, estimated by using non-parametric methods, and the regression coefficients' parametric estimates. In addition, a third ever-present challenge-model selection-can also significantly affect predictions.

To better understand the effect of various levels of truncation and censoring on the predictions, three simulation studies were conducted based on Weibull simulated data with varying amounts of bias. The basic setup in all three simulations was the same. Data sets of sizes $n = 100, 200, 500, 1000, 2000$, and 4000 were generated from a Weibull regression model, a function of one explanatory variable referring to *age*. In the simulation, *age* was uniformly distributed from 22 to 70 years, a range mimicking the actual distribution of worker ages observed in the sample. The simulation was used in the context of retirement, and most of the results were also appropriate for quitting.

The baseline hazard for Weibull distribution with shape α and scale λ is $h_0(t) = \alpha(\lambda)^\alpha t^{\alpha-1}$. Extending this baseline hazard to a hazard from a proportional hazards regression model for *age*, it is simply multiplied by the exponentiated linear predictor, thus shifting the

baseline up or down,

$$\begin{aligned} h(t|age) &= h_0(t)exp(\beta \times age) \\ &= \alpha(\lambda(exp(\beta \times age)^{\frac{1}{\alpha}})^{\alpha} t^{\alpha-1} \\ &= \alpha(\tilde{\lambda})^{\alpha} t^{\alpha-1} \end{aligned} \tag{3.9}$$

where, $\tilde{\lambda} = \lambda(exp(\beta \times age)^{\frac{1}{\alpha}}$.

The survival times T_i are randomly generated from the Weibull distribution with shape parameter $\alpha = 1.5$ and $\tilde{\lambda} = exp(1.5 + 0.025 \times age)^{\frac{1}{\alpha}}$. It follows that $\lambda = e^1$.

The simulations were performed using the `coxreg` and `phreg` functions from the `R`-package `eha` (Brostrm, 2015) for model fitting. Function `coxreg` performs a standard Cox PH regression using the partial likelihood to fit the model. The `phreg` function performs a parametric proportional hazards regression using both Weibull, Extreme value (EV) baselines.

3.5.1 Simulation 1: Right Censoring

The first study focused on understanding right censoring's impact, a significant effect in the turnover data analyzed later because of the many employees that remained active for the entire observation window. For each of the sample sizes above, survival times T_i are simulated from the Weibull distribution as described earlier. Four censoring times C_j are defined as the first, second, third, and fourth (maximum) quartiles of the simulated sample of lifetimes and refer to 75%, 50%, 25% and 0% censoring proportions, respectively. If the i^{th} observation's survival time T_i is below the censoring time (C_i), then the lifetime T_i is observed and the censoring indicator $\delta_i = 1$. When the survival time T_i for i^{th} observation is greater than the censoring time (C_i), then the censoring time C_i is observed and the censoring indicator $\delta_i = 0$.

The results of 100 simulations at each combination of sample size and censoring proportion are shown on the left side of Table 3.1. Column 1 gives the censoring proportion, column 2 indicates the observed number of events before censoring, and columns 3 and 4 show the average β_{age} estimates over 100 simulations for both `coxreg` and `phreg`. The values

in columns 5 and 6 are the average estimates of λ and α from the parametric fit of `phreg`. The simulation results show that censoring proportion and the number of events are two influential factors in estimating the coefficient. The model overestimates the coefficients of age, λ, and α, when the dataset has a high proportion of censoring. For example, when 75% of the data are censored with only 25 events, the estimates for three parameters are 0.028, 4.043, and 1.564, respectively, which are the highest among all the estimates. As the event number increases, the estimates approach the actual value. For example, the estimation of age, λ, and α are close to 0.025, 2.7, and 1.5, respectively, as the number of events exceeds 500.

Table 3.1: Right Censoring and Left Truncation Simulation Statistics

Right Censoring						Left Truncation					
Right Censoring Proportion	Events	β_{coxreg}	β_{phreg}	λ_{phreg}	α_{phreg}	Left Truncation Proportion	Events	β_{coxreg}	β_{phreg}	λ_{phreg}	α_{phreg}
0%	100	0.026	0.027	2.931	1.509	0%	100	0.027	0.027	2.865	1.534
25%	100	0.027	0.027	2.962	1.527	25%	75	0.027	0.027	2.917	1.546
50%	100	0.028	0.028	3.237	1.530	50%	50	0.027	0.027	2.899	1.577
75%	100	0.028	0.028	4.043	1.564	75%	25	0.029	0.029	3.280	1.757
0%	200	0.026	0.026	2.841	1.508	0%	200	0.025	0.025	2.777	1.506
25%	200	0.026	0.026	2.856	1.513	25%	150	0.025	0.025	2.756	1.515
50%	200	0.026	0.026	2.925	1.527	50%	100	0.025	0.025	2.825	1.532
75%	200	0.026	0.026	3.167	1.540	75%	50	0.026	0.026	2.927	1.572
0%	500	0.025	0.025	2.731	1.500	0%	500	0.025	0.025	2.732	1.509
25%	500	0.025	0.025	2.718	1.508	25%	375	0.025	0.025	2.737	1.514
50%	500	0.025	0.025	2.744	1.514	50%	250	0.026	0.026	2.778	1.514
75%	500	0.025	0.025	2.787	1.525	75%	125	0.026	0.026	2.835	1.547
0%	1000	0.025	0.025	2.748	1.509	0%	1000	0.025	0.025	2.710	1.504
25%	1000	0.025	0.025	2.747	1.512	25%	750	0.025	0.025	2.709	1.504
50%	1000	0.025	0.025	2.748	1.514	50%	500	0.025	0.025	2.715	1.506
75%	1000	0.026	0.026	2.844	1.509	75%	250	0.025	0.025	2.694	1.524
0%	2000	0.025	0.025	2.714	1.502	0%	2000	0.025	0.025	2.740	1.503
25%	2000	0.025	0.025	2.713	1.503	25%	1500	0.025	0.025	2.731	1.502
50%	2000	0.025	0.025	2.742	1.500	50%	1000	0.025	0.025	2.724	1.503
75%	2000	0.025	0.025	2.733	1.502	75%	500	0.025	0.025	2.718	1.508
0%	4000	0.025	0.025	2.719	1.504	0%	3999	0.025	0.025	2.720	1.500
25%	4000	0.025	0.025	2.718	1.505	25%	3000	0.025	0.025	2.722	1.501
50%	4000	0.025	0.025	2.724	1.503	50%	2000	0.025	0.025	2.710	1.502
75%	4000	0.025	0.025	2.729	1.513	75%	1000	0.025	0.025	2.703	1.503

3.5.2 Simulation 2: Right Censoring with Staggered Entry Times

To capture more accurately the data sampling scheme's nuances and complexities and to understand right censoring's impact, the above simulation was modified by staggering the entry times. Starting with the Weibull simulated failure times, offset factor S was added

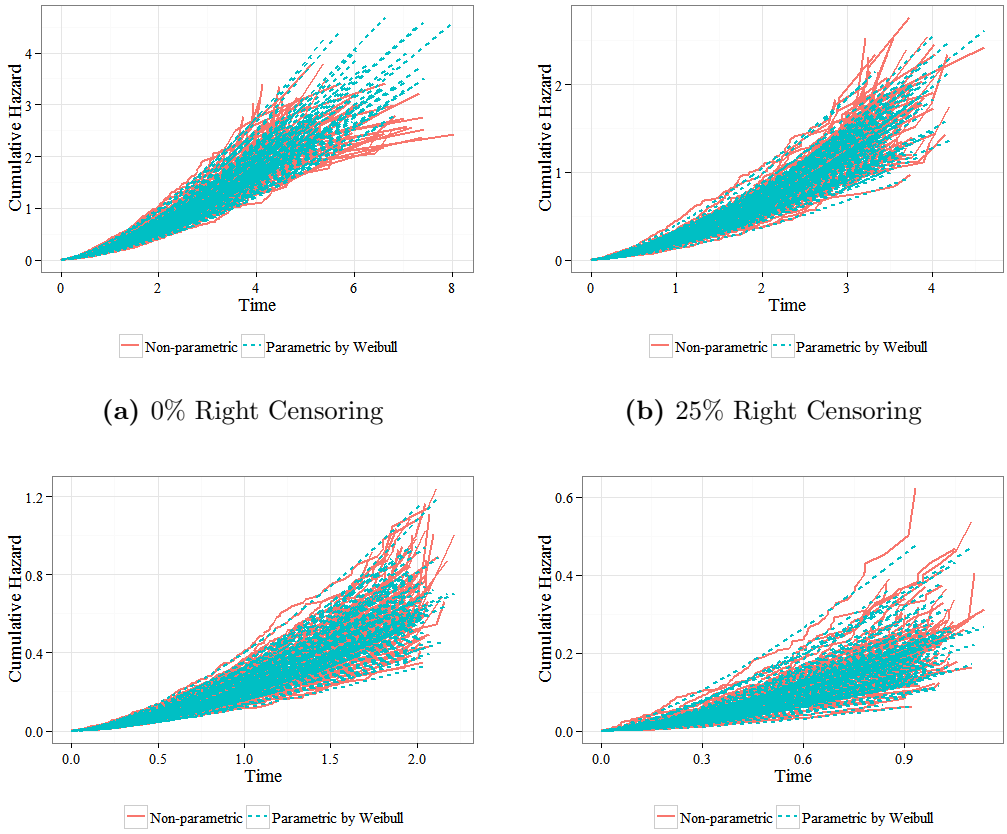
following a uniform distribution from 0 to 10 and representing variation in the employees' starting times within the study window. The event time is equal to the sum of the start point and survival time: $S+T$. The censoring time is a single fixed value that ensures a fixed proportion (25%, 50%, and 75%, respectively) of censored observations. The observation and censoring indicator are then determined as in the first simulation an individual's survival time being $min(C, T_i + S_i) - S_i$. Because some observations started after the cutoff point (censoring time), the sample sizes varied for different censoring proportions. To ensure a constant sample size, 6000 observations were initially simulated, and 400 whose start points occurred before the censoring point were randomly selected.

Table 3.2: Right Censoring Simulation's Results Based on Various Start Time

Censoring Pct.	Events	Variable Estimates			Predicted Events	
		age	λ	α	*coxreg*	*phreg*
0%	400	0.025	2.694	1.508	398.52	400.44
25%	400	0.026	2.802	1.518	394.24	401.72
50%	400	0.026	2.828	1.514	340.73	398.51
75%	400	0.025	2.821	1.518	215.92	400.80

The simulation results are shown in Table 3.2. As discussed above, both a correctly parametric proportional hazards regression model with a Weibull baseline `phreg` and a semi-parametric Cox PH regression `coxreg` were fit in order to evaluate differences in efficiency. The estimation of age and α are close to the true values (0.025 and 1.5) when averaged over 100 simulations. Based on 400 events, the estimates of λ increase from 2.694 with no censoring to over 2.8 when censoring is above 50%. In general, the results indicate that right censoring has little impact on coefficient estimation and that the semi-parametric estimates show efficiency similar to that of fully parametric estimates. However, right censoring does affect the baseline function's estimation for the Cox PH model as shown in columns 6 and 7 of Table 3.2 and Figure 3.2. Panel (a) shows no censoring, and the parametric baseline matches the non-parametric. As time increases and the number at risk initially decreases, the non-parametric estimate deviates from the parametric fit. Panel (b) shows 25% censoring for 100 replications. As the data range decreases, the non-parametric baseline estimate's duration is more restricted, while the parametric fit can be extrapolated with the usual caveats. In panel (d), the 75% censoring level, an increased variability is shown with the diminished range because of the more restrictive censoring time. Note that these results are

indicative of monotonically increasing hazards and may differ if other baseline distributions are considered.



(a) 0% Right Censoring

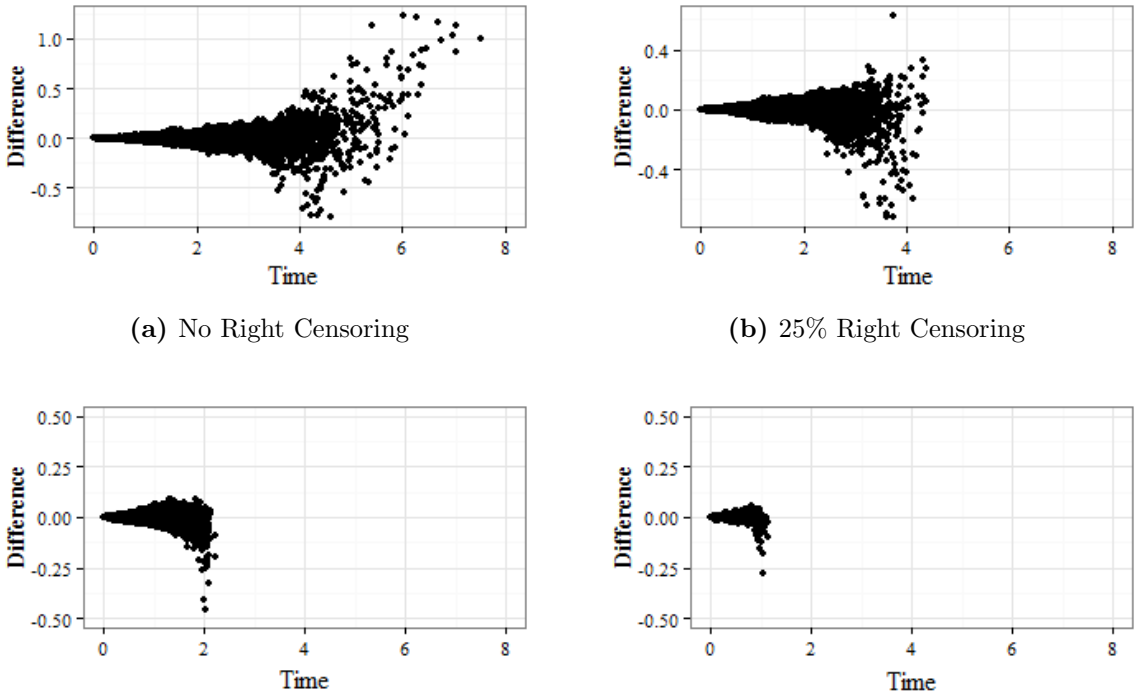
(b) 25% Right Censoring

(c) 50% Right Censoring

(d) 75% Right Censoring

Figure 3.2: Baseline Comparison by Various Censoring

Figure 3.3, panels (a-d) illustrate the differences between parametric and non-parametric estimates of the baseline cumulative hazard function, $H_0(t) = -log(\hat{S}(t))$ at different time points over 100 simulated data sets. Panel (a) indicates that without censoring, overall survival estimates including both baseline and covariates are very similar across parametric and semi-parametric approaches. As lifetime t increases, deviations between the two estimates increase as a result of both the estimates' cumulative nature and the increasing variability of the non-parametric baseline's hazard estimation, i.e., the Breslow estimator (Davison and Hinkley, 1997; Burr, 1994), as the number of observations at risk diminishes over time.



(a) No Right Censoring

(b) 25% Right Censoring

(c) 50% Right Censored

(d) 75% Right Censoring

Figure 3.3: Differences in Estimates of Baseline Cumulative Hazard Functions, Parametric and Non-parametric, for four Levels of Censoring

When censoring is introduced in panel (b), the range of the x axis is restricted. Although this restriction varies slightly across simulations, the non-parametric estimator clearly cannot estimate survival probabilities beyond the maximum observed lifetime because of the lack of a parametric model for the baseline. In the current simulation, this estimate was dictated by the censoring time, which was set to ensure a fixed proportion of censored cases. In real studies, such as the one introduced in Section 3.6, the population and sampling window dictated the maximum observed lifetime. Panels (c) and (d) reiterate both of these factors. As the censoring level increases to 50% and finally 75%, the range of non parametric baseline estimates is further restricted, but the accuracy improves because of the increasing concentration of observations before the censoring time. Because all simulations had 400 observations, 75% censoring ensured 100 events before the censoring point, which is close to 1, thus providing a very accurate baseline estimate and less dispersion than the uncensored

data. Parametric models do not suffer from limitations on the baseline estimate. It is a visual advantage but requires careful model-selection steps, which introduce other challenges.

3.5.3 Simulation 3: Left Truncation

Finally, a simulation was preformed to evaluate the impact of left truncation bias on parameter estimation and prediction. Again, a simulated sample of employment times $T_1, \ldots, T_n$ was generated. For each observation, i, a uniform random variable $S_i \sim U(0, max(T))$ was then generated, representing the simulated employee starting time. The event time is then $R_i = S_i + T_i$. Figure 3.4 provides a histogram of the simulated population of R_i. Four levels of truncation are introduced by shifting the beginning of the sampling window, L from 0 across the quartiles of R . When the i^{th} observation's starting time, S_i, is less than truncation time l_i, the observation starting point was reset to l_i which is the first point at which the employee is observed in the sampling window. If $S_i > l_i$, then the employee was first observed at S_i, which remained the starting point. If $R_i > C$, where C represents the end of the observation window, then the employee was still active at the end of the sampling window and that employee' turnover time was censored. To isolate the impact of truncation bias, no censored values were generated in this study.

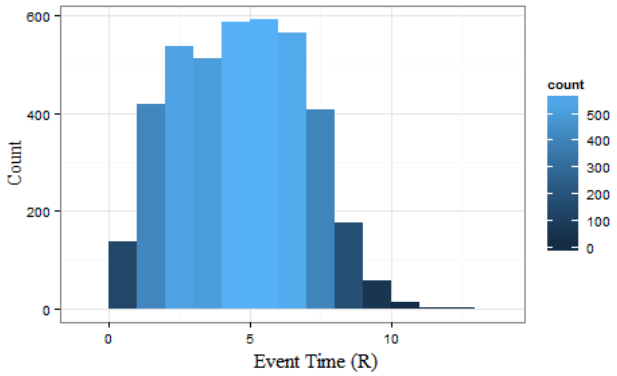
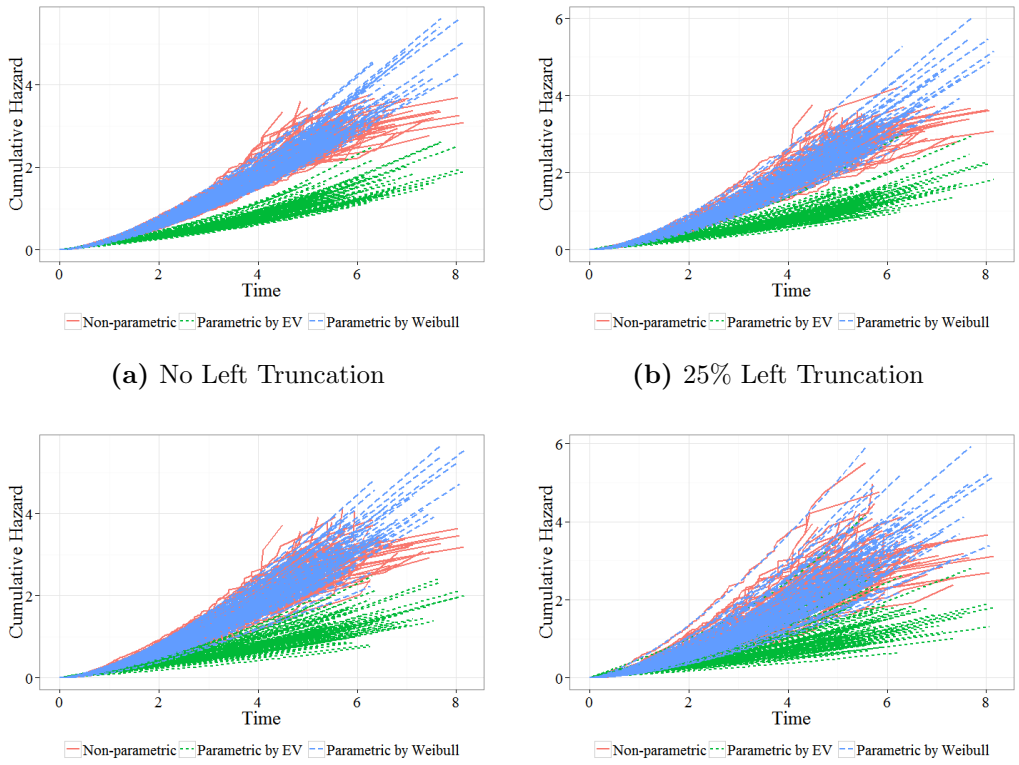


Figure 3.4: Histogram of Simulated Lifetimes

Sample size, censoring proportions, and results follow the same protocol given in Section 3.5.1. The simulation's results are shown on the right side of Table 3.1. As before, the values shown represent average parameter estimates over 100 replications for the Cox PH model and

a Weibull PH model. The general pattern suggests that as truncation proportion increases for a fixed number of events, the parameter estimates become slightly biased. In the Cox PH case when the sample size is 100, the coefficient for age increases from .027 to .029 as the truncation increases from 0% to 75%. The scale parameter for the parametric model, `phreg`, also increases with truncation percentage. For samples of size 100, with $\lambda = e^1 \approx 2.718$ the average estimate increases from 2.865 with no truncation to 3.280 with 75% truncation. Estimates for $\alpha = 1.5$ increase from 1.534 to 1.757. Such effects disappear as the number of events increase.



(a) No Left Truncation

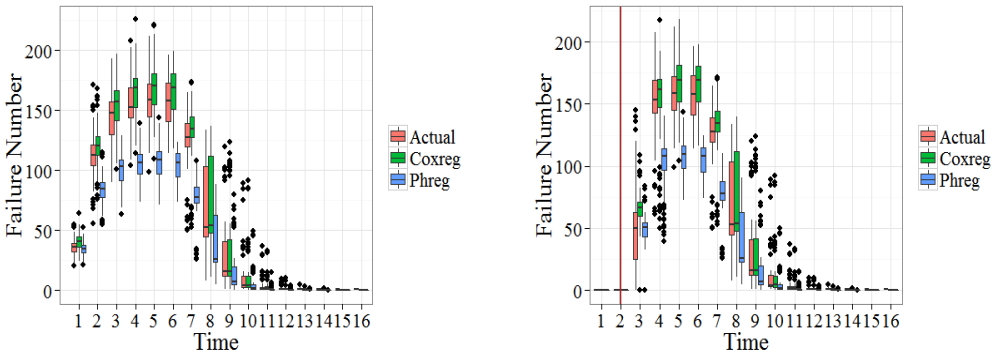
(b) 25% Left Truncation

(c) 50% Left Truncation

(d) 75% Left Truncation

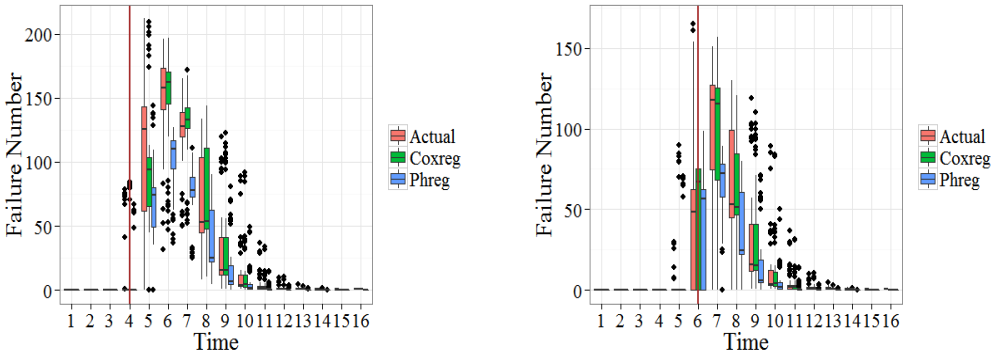
Figure 3.5: Left Truncation Simulation Predictions: Comparison of Actual vs. Cox PH and Parametric PH with EV Baseline

Figure 3.5 compares the baseline's non-parametric estimates from the Cox PH model with two parametric proportional hazards fits estimated using `eha` (Brostrm, 2015). The first parametric fit is properly specified and assumes a Weibull distribution as in the prior simulations. The second fit assumes that data follow a type I extreme value distribution.



(a) No Left Truncation

(b) 25% Left Truncation



(c) 50% Left Truncation

(d) 75% Left Truncation

Figure 3.6: Left Truncation Simulation Results: Comparison of Actual vs. Cox PH and Parametric PH with EV Baseline Predicted Failure Number

Unlike the simulation described in Section 3.5.2, the non-parametric baseline estimates' duration is not limited by the four left truncation proportions. Both Weibull and Cox PH fits overlap with the Cox baselines' variability showing increased variability with the amount of truncation. EV based fits all significantly underestimate the cumulative hazard indicating a potential risk of parametric modeling. However, it is interesting to note that the EV fits converge slightly with increasing amounts of truncation as a result of extreme-value theory and the theory of exceedances. More information about this phenomenon can be found in Coles et al. (2001).

Figure 3.6 focuses on the number of events that the models predict. Here, 2000 events were generated using this section's protocol. Both parametric and nonparametric models were fit to the data, and this procedure was repeated 100 times. The red vertical line in Figure 3.6 is the average left-truncation time across the simulations for each truncation

percentage. Using each model, the expected number of failures was then predicted for each unit's time interval and compared to the observed number of failures. Boxplots indicate the range of observed values across the 100 replications. The non-parametric Cox PH model tended to slightly overestimate the number of events in each interval, indicating that left truncation did not affect baseline estimation. The convention used to compute the survival probability over each interval explains the overestimation. The baseline survival function is stepwise decreasing, and probabilities are based on the estimate at the most recent previous failure. The baseline slightly overestimates the failure for each individual, and in aggregate produces the observed overestimates. Matching parametric estimates are generated using a PH model with extreme value (EV) baseline instead of the true Weibull model. The EV fit's underestimation of the baselines shown in Figure 3.5 leads to drastic under-predictions of the number of events.

In the situations considered, if sample sizes are large enough, particularly if they exceed 250 events, all parameter estimates show very little bias under both parametric and semi-parametric models. If samples are small, regression coefficients are slightly overestimated in both types of models. For data sets with small numbers of events because of heavy right censoring, parametric models tend to overestimate both shape and scale, leading to an underestimated baseline. Although right censoring's impact on the non-parametric baseline is not directly quantified, Figure 3.3 suggests that baseline estimates are often smaller than the parametric versions and become more variable near the maximum lifetime. Simulations show similar impacts of left truncation on covariate and baseline effects. A major difference is that the non-parametric baseline's length is potentially less limited because of the lack of censoring time. This may not be relevant in real world data, which has both effects.

Although this study has more than 50% right censoring and an unknown proportion of left truncated cases, it still has more than 1000 retirement events, many with long duration (around 50 years). Based on the simulation, the Cox model is expected to provide very accurate estimates of covariate effects in this case and highly accurate baseline estimates for most employees, particularly those under the age of 70.

3.6 Results and Analysis

Constructing predictive models for turnover involves several factors. Among the variables available for analysis, the set that offers the maximum predictive power must be chosen (i.e., a model that includes variables providing the best possible predictions on out-of-sample testing data sets and not simply on data used to train the model). The baseline hazard estimate's potential strengths and weaknesses and the impact on prediction accuracy must be evaluated. A bootstrap is a useful tool for assessing uncertainty in model predictions. From an administrative perspective, it is also useful to contemplate the predictors' ERIP implications found to have an impact on prediction.

3.6.1 Descriptive Analysis

As described in Section 3.3, the current data provides five demographic and career history variables: PR (hourly, weekly, or monthly payroll); GENDER (M, F); DIV (ten divisions); OC (crafts, engineers, general administrative, laborers, managers, administrative, operators, scientists, technicians); AGEH; and YCSH. Table 3.3 provides marginal counts of the number of workers in the sample within each category. The occupation codes include four large categories (C, E, M, &P) with over 1000 employees observed throughout the data sample; four medium-sized groups with 500-700 employees (G, L, R, &T); and one small group, S, with 208 employees. Payroll data shows that the largest group is paid monthly, followed by hourly, and weekly. In terms of gender, approximately 72% of employees are male. Finally, divisions, while not fixed over the employee's life, are distributed in a similar fashion to occupational codes with four larger groups and several of smaller groups.

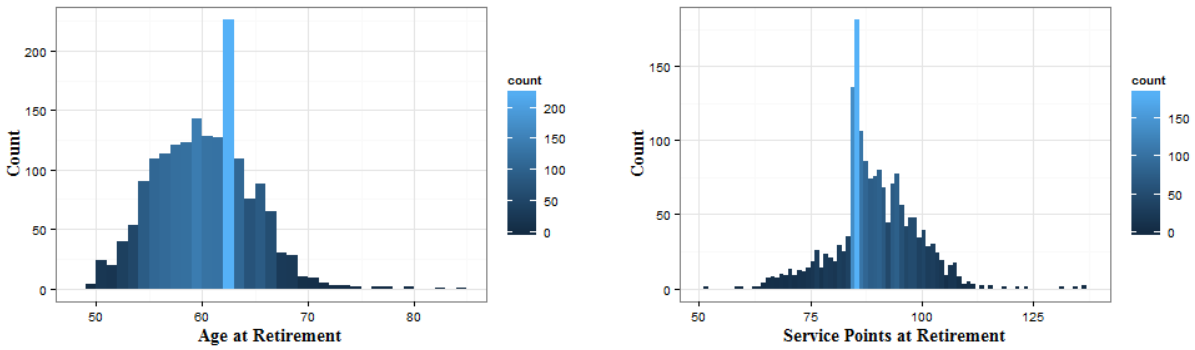
Beyond these demographic and career factors, the models also include behavioral variables derived from policy requirements for retirement and early retirement incentives occuring throughout the observation period. Histograms of retirement age and accumulated pension points are shown in Figure 3.7. The histogram of retirement age is right skewed and shows an anomalous spike at age 62, which is the mode. Also, 79% of the employees who retired at 62 also reached or exceeded 85 points.

Table 3.3: Descriptive Statistics

Variable	Count	N %	Variable	Count	N %
OC			**GENDER**		
C	1295	16.0%	F	2296	28.4%
E	1361	16.8%	M	5802	71.6%
G	574	7.1%	**DIV**		
L	613	7.6%	Div1	1542	19.0%
M	1178	14.5%	Div2	751	9.3%
P	1621	20.0%	Div3	1042	12.9%
R	595	7.3%	Div4	369	4.6%
S	208	2.6%	Div5	398	4.9%
T	652	8.1%	Div6	1199	14.8%
Missing	1	0.0%	Div7	302	3.8%
PR			Div8	823	10.2%
Hourly	2503	30.9%	Div9	404	5.0%
Monthly	4369	54.0%	Div10	1268	15.7%
Weekly	1226	15.1%			

Table 3.4: Descriptive Statistics 2

	Count	Mean	Median	Mode	Minimum	Maximum	Std. Deviation
Age at Retire	1757	59.72	60.00	62.00	49.00	84.00	4.56
Years of Service at Retire	1757	29.72	30.90	30.06	0.05	55.68	7.74
Points at Retire	1757	89.44	88.67	85.47	51.05	136.66	9.15



(a) **(b)**

Figure 3.7: Histogram of Age and Point at Retire

The average retirement age is 59.72, demonstrating that many individuals retire before 62 and most before age 70. In terms of points (i.e., the sum of years of service plus current age) accrued at time of retirement, an irregular distribution is shown with the vast majority retiring with total points ranging from 85 to 100. Relatively few employees elect to take a reduced pension and retire with diminished benefits with points below 85. Again, 85 points

is the mode, indicating that retiring immediately after becoming fully vested in the pension plan is a popular choice.

3.6.2 Retirement Models without External Economic Variables

Section 3.4.7 discussed two potential response variables suitable for time-until-turnover models: age and years of service from hire. Because of the existence of time-varying variables and the need for a baseline estimate to facilitate predictions, age at retirement is chosen arranging the data in a counting process format (SAS only estimates the baseline if the counting process formulation is used).

Table 3.5: Retirement Models' Assessment

No.	Model	LR	AIC	SBC	Pred. MAPE	Holdout MAPE	Pred. G^2	Holdout G^2
1	DIV GENDER PR OC YCSH AGEH	1194.30	19271.3	19377.3	39.44	56.78	381.77	85.19
2	DIV OC YCSH AGEH	1193.80	19269.8	19370.5	39.45	56.84	381.92	85.29
3	DIV ERIP YCSH AGEH	1451.60	18998	19061.6	25.91	15.51	128.75	2.64
4	DIV ERIP YCSH AGEH OC	1469.92	18995.7	19101.7	25.91	15.24	129.47	2.56
5	DIV ERIP YCSH AGEH P85	1826.95	18624.6	18693.6	25.59	19.04	109.17	3.40
6	DIV ERIP YCSH AGEH P85 P85*A65	1873.69	18579.9	18654.1	25.38	7.97	111.55	0.79
7	DIV ERIP YCSH AGEH P85 P85*A65 P85*ERIP	1881.02	18574.6	18654.0	25.42	4.20	112.27	0.81
8	Time series	N/A	N/A	N/A	11.17	34.38	32.10	8.54

Extensive model selection using a variety of metrics including log-likelihood, AIC, BIC, and out-of-sample predictive scoring (MAPE and G^2) was then applied to identify key predictive factors in the model as shown in Table 3.5. The first four models listed are all standard Cox PH models with various sets of explanatory variables. Among these four, the third model offers the lowest BIC value (19,061), while the fourth has the lowest AIC (18,996) indicating these two models are the best of this subset. Overall, both models perform almost identically on MAPE and holdout MAPE, which is error when the model is used on the out-of-sample data from 2011 and 2012. G^2 and holdout G^2 were also nearly identical. From this perspective, the variable OC seems to provide very little predictive impact.

Models 5-7 were built on model 3 and added various interaction terms. Again, models 6 and 7 are roughly equivalent to the more complex model having a smaller AIC and holdout MAPE, and a very slightly higher predicted G^2 and predicted MAPE. In the end, the more complex model 7 was selected and explored because of interest in interpreting the interacting model coefficients.

Stratification is another modeling technique that creates a separate baseline for of a categorical predictor's levels. Creating multiple baselines for subgroups makes the model significantly more complex and can hugely reduce log likelihood and traditional model fit. However, this flexibility can also lead to over-fitting and very poor predictive and holdout fits if the stratification variables are not properly chosen or data is lacking at some levels. This methodology was tested but rejected because of the poor predictive performance in the holdout sample.

Finally, for comparison purposes, a time series prediction model was included as discussed in Chapter 2. This model was fit to monthly aggregate retirement counts from November 2000 to December 2010 based on the same data used here. The model produced an aggregate forecast by month, leading to a very accurate MAPE on the training data but much poorer out-of-sample performance. It did not provide accurate forecasts by occupational code (OC) because of the small individual samples for each subgroup. Furthermore, including other explanatory factors' effects in this model was impossible.

Excluding aggregate economic factors, the optimal modelling variables for predicting age at retirement include DIV, YCSH, and AGEH. In addition, based on the understanding of the retirement program's covenants and parameters, several additional variables increasing the model's predictive power were found. These variables include ERIP, P85, A65*P85, and ERIP*P85. A65*P85 moderates the P85 effect's impact after the individual has exceeded 65 years of age. ERIP*P85 moderates the P85 effect's impact while the ERIP is in place. These variables were introduced in Section 3.3.

Table 3.6 describes the fit parameters and hazard ratios. As noted above, gender, occupational code and payroll category were not statistically significant predictors, indicating that employees' gender, job types, and payroll status are not associated with the choice of retirement age conditional on the other variables in the model. P85 is an indicator that a

Table 3.6: Parameter Estimates for Retirement Models

		Model W/O External Variable		Model with Real Earnings	
Parameter	Label	Parameter Est. (Std. Error)	Hazard Ratio	Parameter Est. (Std. Error)	Hazard Ratio
DIV	Div2	-0.965 (0.179)***	0.381	-0.969 (0.177)***	0.380
DIV	Div3	-0.241 (0.112)*	0.786	-0.242 (0.111)*	0.785
DIV	Div4	0.078 (0.195)	1.081	0.013 (0.195)	1.013
DIV	Div5	-0.131 (0.19)	0.877	-0.216 (0.190)	0.806
DIV	Div6	2.136 (0.095)***	8.463	2.261 (0.093)***	9.594
DIV	Div7	2.435 (0.129)***	11.418	2.515 (0.128)	12.363
DIV	Div8	0.864 (0.106)***	2.373	0.855 (0.106)***	2.352
DIV	Div9	-3.023 (0.581)***	0.049	-2.731 (0.504)***	0.065
DIV	Div10	0.793 (0.093)***	2.211	0.726 (0.093)***	2.068
YCSH	1	0.019 (0.004)***	1.019	0.023 (0.004)***	1.023
ERIP		0.859 (0.169)***	.	0.489 (0.133)***	.
AGEH		-0.172 (0.013)***	0.842	-0.193 (0.014)***	0.825
P85	1	1.435 (0.091)***	.	0.756 (0.073)***	.
P85*A65	1	-1.610 (0.206)***	.	-1.019 (0.197)***	.
ERIP*P85	1	0.469 (0.179)**	.	0.506 (0.141)***	.
Real Earnings				0.013 (0.001)***	1.013

[1] * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$.

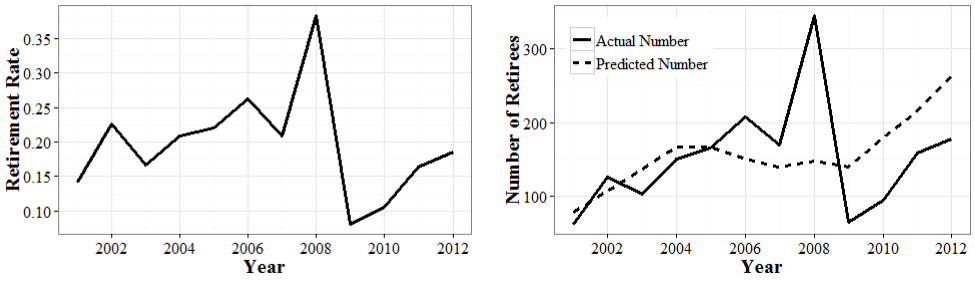
person is eligible for maximum retirement benefits. Naturally, this eligibility has a strong impact on the probability that a person will retire. From a quantitative perspective, with all other factors held constant, for those achieving 85 points before the age of 65, the hazard ratio is $e^{1.44} = 4.22$ meaning that the hazard of retirement becomes 4.22 times more likely. While not surprising, this quantification is important in predicting individual and aggregate retirement and reflects the modal spike observed in the histogram in Figure 3.7b.

An alternative eligibility criterion for retirement occurs when individuals exceed age 65; thus, the hazard increases at this point in an employee's career. Because the response variable in the model is age, this 65-year effect's impact is included in the baseline hazard, which should increase after this point. Figure 3.9 shows the baseline survival and cumulative hazard function for a standard case. Independent of the P85 effect, a further steep increases in the cumulative hazard/decrease in survival between age 62 and 65. By including an interaction between the indicators of age greater than 65 and points greater than 85 (A65*P85), how the impact of reaching 85 points changes when a person exceeds regular retirement age can be estimated. In this case, the interaction term is estimated at -1.61, indicating a diminishing effect on the P85 criterion to $e^{1.44-1.61} = 0.84$. In addition, as can be seen from the cumulative hazard's trend, the baseline hazard seems to return to a lower level. Hence, those employees

who that exceed both criteria actually have a reduced hazard of retiring over those who have met only the age 65 criterion. If the individual remains on the job after meeting both criteria seemingly indicates a diminished intention to retire.

According to the model, an employee's age at time of hire and their years of service at the time of hiring can also influence retirement. The coefficient estimate for age is -0.17. The reference age is 45.49 (i.e.,the hazard ratio for retirement of an employee who started working at age 46.49 is $e^{-.017} = .84$) indicating a 16% drop in hazard for each additional year later that an employee started. The employee's survival probability at any time (t) can be computed as $S(t)^{1.19} = (S(t)^{e^{0.17}})$ when age at hired is one year below 45.49, where $S(t)$ is the baseline survival probability for a reference employee of average age at the time of hiring. Moving in the other direction, an employee's survival probability is $S(t)^{0.84} = (S(t)^{e^{-0.17}})$ for a one-year increase beyond 45.49 in the employee's starting age. Together, the estimates of age and years of service imply that at any given retirement age, the employee who starts earlier is more likely to retire because of having more years of service and being closer to vesting full benefits (85 points) than an equivalent employee who started working at an older age. Similarly, the employee's years of service at hire show a positive estimate (0.019) with a hazard ratio (1.019) indicating that each year of service at hire beyond the population average of 2.75 is associated with an approximately 2% increase in the hazard of retirement. This increase leads to a survival probability $S(t)^{0.98}$ for a one-year decrease in the reference years of service at time of hiring. On the other hand, the survival probability is $S(t)^{1.019}$ for an employee with one year of service above average at the time of hire. Together, age at hire and YCSH effects reflect the intuitive fact that, all else being equal, an employee who has more years of service and therefore is closer to full vesting is more likely to retire. What is non-intuitive about this finding is that while one might suspect that the effects should be of similar magnitude, the effect of one year difference in age actually seems to have about eight times the impact that one year of previous service does on the hazard of retirement.

In the fiscal year 2008, the employer in this study created a temporary early retirement incentive program (ERIP). The response window for this option was three months; however, other details are unknown. To deal with the increased retirement level during this period, a time-dependent indicator variable was included for each employee who indicated their age



(a) Retirement Rate **(b)** Prediction w/o Variable ERIP

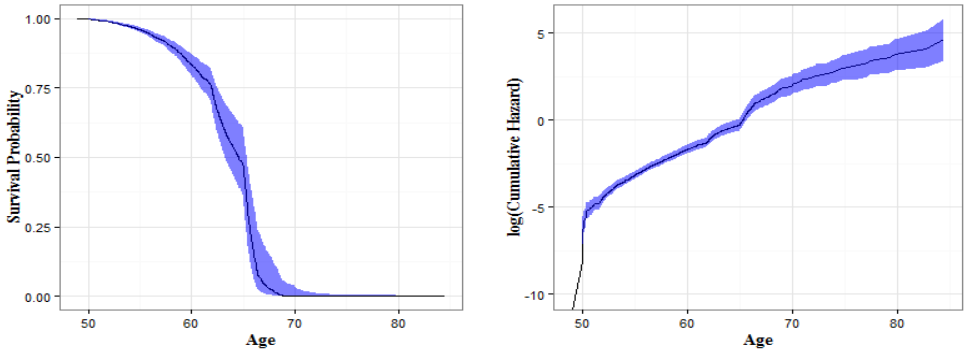
Figure 3.8: Retirement Rate and Actual vs. Forecast Plot by Preliminary Model without ERIP

when this program was in effect. This indicator's coefficient was 0.86, leading to a hazard ratio of $e^{.86} = 2.21$, which indicates that, on average, an individual's hazard of retirement increased almost 2.2 times during this period. If more information were known about this ERIP's requirements or targets, a more case-specific estimate might be possible. The ERIP's effect is considerable as indicated by the huge uptick in events in 2008 (see Figure 3.8). Not including this one-time effect in the model might bias the other estimates' parameters considerably.

As an additional step, the ERIP effect was tested on the employees eligible for a pension (i.e., employees with points greater than 85). After adding an interaction term between ERIP and the indicator that a person exceeded 85 points, the hazard ratio for the ERIP increases substantially from 2.21 to 15.85, more than seven times the basic ERIP effect. In contrast, employees who were eligible for only partial retirement are not affected by ERIP because an interaction between ERIP and indicator variables that employees achieve only 65 or 75 points, were not statistically significant.

The DIV variable was also a significant predictor. For analysis, the baseline level was chosen arbitrarily as division 1 so that the baseline determined its hazard rate. Relative to this baseline, divisions 6 and 7 have very high hazard ratios (8.363 and 11.405, respectively), indicating-with other factors being equal-that the employees in division 6 and 7 are much more likely to retire at any age than those in division 1. Conversely, division 9 has a hazard ratio of $e^{-3.023} = .049$, indicating that individuals within this group have 1/20 the hazard of group 1. This finding may indicate that the division is new and contains younger employees.

In general, differences in retirement rates could be a result of differences in age demographics, departmental and job functions, or departmental leadership.



(a) Survival Function **(b)** Cumulative Hazard

Figure 3.9: Baselines with 95% Confidence Intervals

The baseline survival function and log cumulative hazard function are shown in Figure 3.9. The survival probability is 1 before age 49 as shown in Figure 3.9a, indicating that no employees retire before this age. The survival probability starts to slowly decrease from age 50 to 62. By age 62, the survival probability has decreased by nearly 25%, indicating that 75% of employees retire at an age greater than 62 years, assuming that they are at average or baseline levels for other factors included in the model. The survival function's slope decreases sharply at this point, indicating the increased retirement rate for workers between age 62 and 65. After age 65, the probability drops off even more because most of the remaining population retires by age 68 or 69. Accompanying the survival function is the cumulative hazard ratio's log. Again, the steep rise in the cumulative hazard between age 62 and 65 indicates the increased retirement activity during this period. Afterward, the cumulative hazard levels off, indicating a drop in the instantaneous hazard rate at these future points.

Aggregate predictions for employee retirement are shown in Figure 3.10 and Tables 3.7. The model predictions capture not only the fluctuations in actual retirement but also the peak year 2008 when the ERIP was introduced. The out-of-sample predictions for holdout years (2011 and 2012) are very close to the actual number, indicating that the model performs well on both the training and holdout samples. Besides predicting overall retirement, the

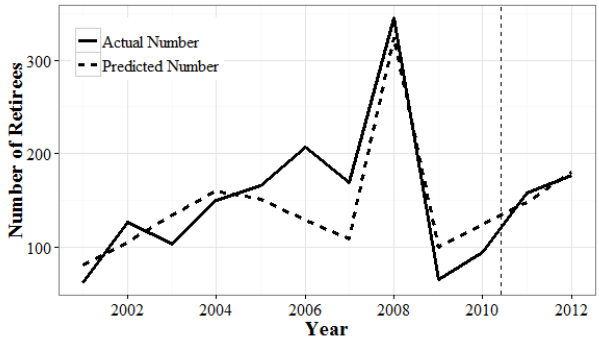


Figure 3.10: Actual vs. Forecast Retirement Number for the Model without Financial Index

Table 3.7: Predictions Comparisons by Occupational Code (OC)

	RE w/o External		RE w. External		VQ w/o External		VQ w. External	
	Training	Holdout	Training	Holdout	Training	Holdout	Training	Holdout
Crafts	31.3[1](34.8)[2]	23.0 (32.5)	27.4 (34.8)	26.9 (32.5)	1.5 (1.7)	1.0 (0.5)	1.5 (1.7)	1.0 (0.5)
Engineers	17.6 (18.9)	14.5 (24.0)	16.9 (18.9)	18.4 (24.0)	16.8 (20.4)	8.2 (7.5)	16.9 (20.4)	7.4 (7.5)
Gen. Admin.	7.9 (8.6)	12.5 (9.5)	7.4 (8.6)	16.4 (9.5)	3.1 (3.4)	2.9 (5.5)	3.0 (3.4)	2.7 (5.5)
Laborers	9.2 (8.1)	10.5 (9.5)	8.5 (8.1)	10.9 (9.5)	2.1 (2.4)	1.1 (1.0)	2.1 (2.4)	1.0 (1.0)
Managers	23.3 (27.0)	37.5 (36.5)	20.7 (27.0)	45.6 (36.5)	6.6 (7.1)	3.3 (3.5)	6.7 (7.1)	3.1 (3.5)
Prof. Admin.	27.9 (28.3)	44.5 (35.0)	25.2 (28.3)	57.0 (35.0)	11.3 (12.4)	9.9 (8.5)	11.4 (12.3)	9.6 (8.5)
Operators	12.2 (11.9)	10.5 (4.5)	10.7 (11.9)	12.3 (4.5)	1.2 (1.3)	0.6 (0)	1.2 (1.3)	0.6 (0)
Scientists	2.8 (3.0)	2.5 (1.5)	2.7 (3.0)	2.9 (1.5)	1.5 (1.6)	1.1 (2.0)	1.5 (1.6)	1.0 (2.0)
Technicians	8.8 (8.0)	8.5 (14.5)	8.0 (8.0)	11.5 (14.5)	2.8 (3.2)	1.4 (1.5)	2.8 (3.1)	1.3 (1.5)

[1] The number before the parentheses is average predicted number of events.
[2] The number in parentheses is average of actual events.
[3] Training period: Jan. 2001 - Dec. 2010, and testing period: Jan. 2011 - Dec. 2013

model can also predict by category. Table 3.7 columns 1 and 2 show the average actual retirement events (in parentheses) and average yearly predictions by occupational code. These predictions are computed by summing individuals' retirement probabilities by job classification then by year and then averaging. Training (2000-2010) and holdout (2011-2012) periods are reported in columns 1 and 2, respectively. Within the training sample, we see close agreement on average between predictions observed across all job classifications. This close agreement reflects the model's effectiveness but also the fact that this is the average over a longer period and estimates the same data used to train the model. Holdout sample predictions are much more variable, indicating that the estimates are a forecast and are in the shorter averaging period. Nevertheless, some large categories show fairly effective predictions in areas such as managers, laborers, and general administration. Significant

under-predictions are observed in crafts, technicians, and engineers; but such forecasts still provide a useful baseline for managers. In contrast, significant over-predictions are observed in professional administration and the operator category. Although both deviations may lead to significant over-hiring, only the operator category was significantly over-predicted. Note that the overestimation in the operator category matches the underestimation in the technician category and that the model may not distinguish between these categories effectively.

3.6.3 Retirement Models including External Economic Variables

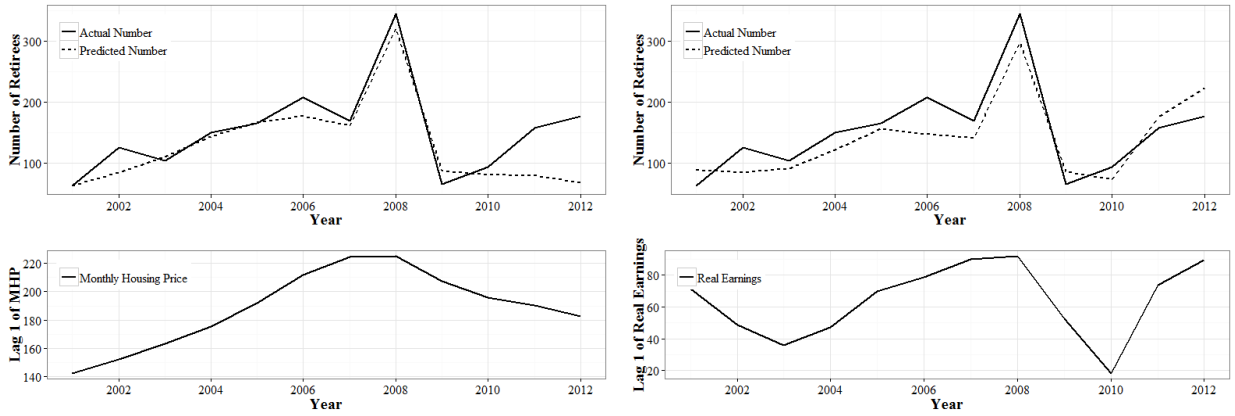
The external economy's impact on retirement decision-making is a topic of considerable interest. To explore this effects, the lagged versions of several economic factors were included using the counting process data formulation with calendar-year based intervals to set up and test the impact on retirement. Parameter estimates for common effects across models with and without economic variables remain constant, as shown in Table 3.6.

Table 3.8: Economic Index Test Statistics for Retirement

Economic Inidcator	χ^2	**P-value**	**Hazard ratio**	**MAPE**	G^2
Without Econmic indicator[1]				21.31	147.43
MHP	129.614	$< .001$	1.020	17.84	220.65
SAMHP	129.516	$< .001$	1.020	17.86	221.66
SEMHP	68.055	$< .001$	1.030	23.37	178.38
SESAMHP	67.871	$< .001$	1.030	23.40	179.87
S&P500	13.319	$< .001$	1.001	20.79	129.83
Dividend	1.045	0.307	1.015	22.63	150.03
Earnings	84.895	$< .001$	1.016	21.40	105.06
Consumer Price Index	5.404	0.020	1.013	21.36	133.57
Real Price	5.522	0.019	1.000	21.22	138.72
Real Dividend	1.925	0.165	1.022	23.39	154.84
Real Earnings	80.358	$< .001$	1.013	20.33	95.02
Long Interest Rate	1.539	0.215	1.082	22.03	149.01
Unemployment Rate	32.212	$< .001$	0.849	25.08	179.49
P/E 10	0.041	0.839	0.998	21.31	147.97
Wilshire5000	22.392	$< .001$	1.028	20.83	121.58

[1] it is the selected model without economic indicator.

As shown in Table 3.8, among financial indices tested, S&P500, REAL EARNINGS, and WILSHIRE5000 were statistically significant and also improved the model forecast leading to



(a) Monthly Housing Price **(b)** Real Earnings

Figure 3.11: Financial Indices and Retirement Predicting Plot

lower MAPE and G^2. Although the REAL PRICES index was also statistically significant, its coefficient estimate of 0.0004 led to a hazard ratio of 1 indicating little if any practical impact. As shown in Table 3.8, the REAL EARNINGS index is the most important factor among all the indicators, leading to the lowest G^2 value and showing the strongest impact on retirement behavior. Figure 3.11b plots retirement fluctuation against a one-year lag of Real Earnings. Two highly correlated equity financial indices, S&P500 and WILSHIRE5000, were also statistically significant with hazard ratios > 1 indicating that this organization's employees are more likely to retire when the stock market is strong.

Unadjusted Monthly Housing Price (MHP) is another influential index, and its inclusion in the model resulted in the lowest MAPE value. Fluctuations in the lag also correlate strongly with the retirement number as shown in Figure 3.11a. However, the retirement number does not decrease coinciding with the decreased MHP after the 2008 financial crisis. Returning to Table 3.7 columns 4 and 5, most of the holdout predictions were inferior to the model without external variables, probably because of the increased deviation between MHP and retirement in the post-2009 period.

Table 3.9 shows yearly predictions for retirement models along with bootstrapped confidence intervals. The third column shows the model predictions based on the fit using the Cox PH model. The bootstrap bias estimate indicates the difference between the predictions and the actual mean values. The bootstrap standard error was estimated using 10,000

Table 3.9: Retirement Models' Bootstrapping Results

		RE w/o Real Earnings				Re w/ Real Earnings			
Year	Actual	Pred.	Bias	Std. Error	$Var\sum\hat{Y}_i$	Pred.	Bias	Std. Error	$Var\sum\hat{Y}_i$
2001	62	80	-0.0524	4.40	91.03	85	-0.2888	6.12	113.40
2002	126	105	0.1736	4.62	112.38	82	0.3079	4.79	96.58
2003	103	134	0.0480	4.62	129.76	90	0.3889	6.10	115.29
2004	150	160	0.2538	4.71	144.18	122	0.4728	5.22	128.09
2005	165	151	0.4936	3.97	130.68	156	0.3703	4.99	143.42
2006	207	129	0.3080	3.59	112.81	146	0.1932	5.55	143.24
2007	169	109	0.2762	3.44	102.96	140	0.1213	7.18	165.28
2008	345	324	0.9460	15.70	472.33	296	0.7722	15.54	457.43
2009	65	99	0.1832	3.35	100.11	86	0.2380	3.54	91.60
2010	94	124	0.2790	3.69	122.31	71	0.5767	4.88	90.09
2011	158	148	0.3432	3.86	139.52	178	0.2685	4.48	170.08
2012	177	180	0.4558	4.41	156.04	226	0.6810	6.95	229.06

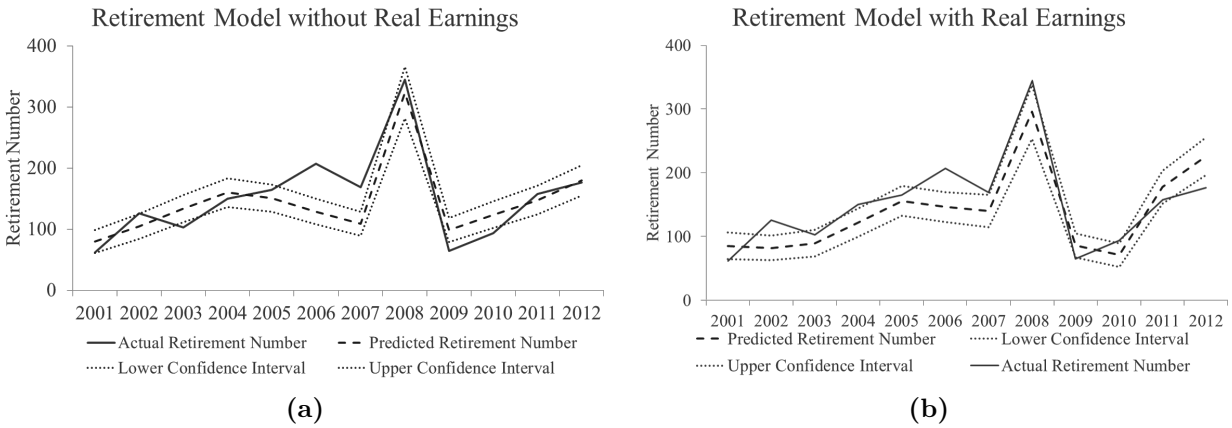
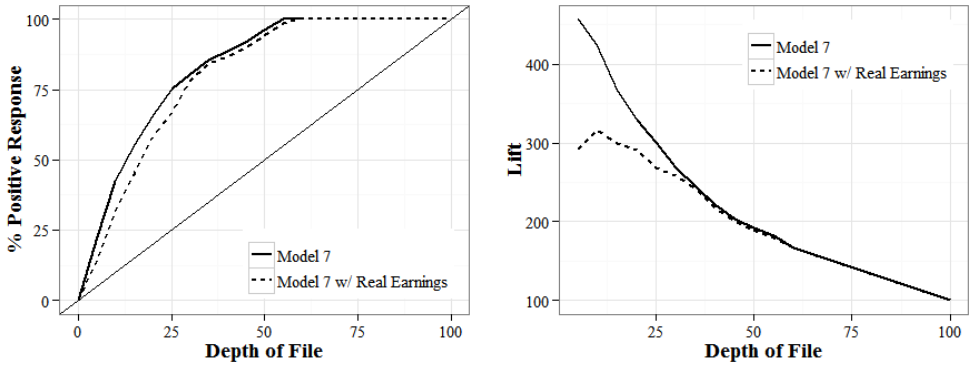


Figure 3.12: Retirement Models' Forecast with 95% Confidence Intervals

bootstrap replications. As shown in Table 3.9, the standard errors range from 3.6 to 15.7 for the retirement model. Although some actual retiring numbers were not included in the 95% confidence intervals, they are close to the predicted confidence intervals. Also, the bootstrap bias estimates are less than 1, indicating the predicted results generated by Cox PH model are nearly unbiased. The retirement models' prediction confidence intervals are shown in Figure 3.12.

Another perspective on the survival models' predictive effectiveness is based on looking at lift and gains plots. Figure 3.13 compares the optimal retirement model (model 7 from Table 3.5) against the same model with the external variable Real Earnings added. The gains plot on the left compares the ranked model fit probabilities (Depth of File) against the cumulative % of positive response. Ideally, if the model fits well, a large proportion of retirees will be

captured in the first percentiles. For example, as shown in Figure 3.13, the 25% of employees predicted most likely to retire contain close to 75% of the actual retirees in 2011; furthermore, the top 50% contain about 95% of the retirees in observed in that year. This figure also compares models with and without Real Earnings to a random-ordering model. The random-ordering model plotted on the diagonal is a worst-case reference, which essentially produces random predictions. In this baseline case, only 10% of retirees are gained for each additional 10% of employees included. Results indicate that the model with Real Earnings included predicts well but is not as effective as the model without those predictions. In either case, the model provides meaningful and accurate predictions of retirement propensity. Figure 3.13b



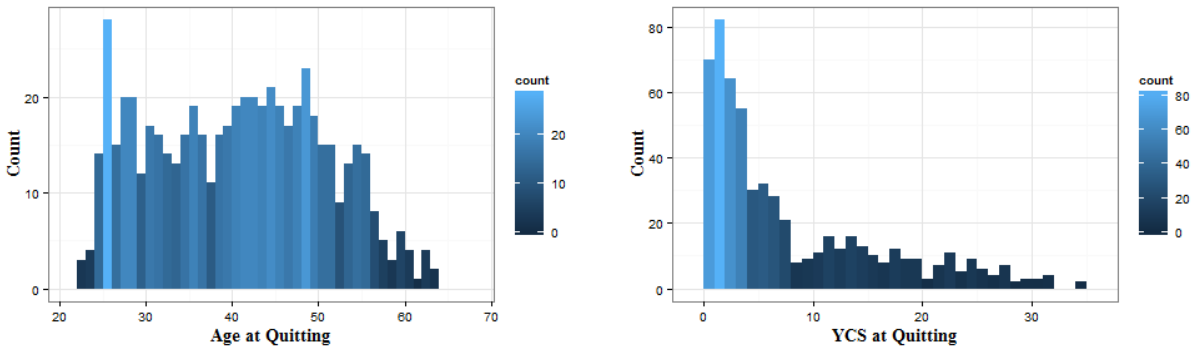
(a) Cumulative Gain Chart **(b)** Cumulative Lift Chart

Figure 3.13: Retirement Models: Gain and Lift Charts of Predictive Efficiency for 2011 Holdout Data

provides another model evaluation perspective, indicating for a fixed sample percentage, how much lift or additional targets, above the baseline percentage is observed in the sample. For example, suppose that 3% of the employees retired in 2011. If a sample of 10% of the employees with highest predicted probabilities is considered from model 7, then the lift plot indicates that about four times as many retirees, or about 12% of this sample, will appear in this group. The model with an external variable performs slightly worse in the first decile, identifying only three times as many retirees or about 9% of the sample. In either case, the model does a reasonable job predicting retirement probability at the individual level (Kuhn and Johnson, 2013).

3.6.4 Voluntary Quitting Models without External Economic Variables

The second area of interest in analyzing turnover is voluntary quitting. Using the same methodology and set of variables applied to retirement modelling, a predictive model for quitting is constructed. The challenge of modelling quitting in the current context is that the data has approximately 600 out of 8000 employees quitting during the 10-year study window, resulting in a very high proportion of censored data. With this data density, the model cannot generate a smooth baseline to achieve a good forecasting model using age as the dependent variable because employees quit across such a wide range of ages (20 to 64); see Figure 3.14a and Table 3.10. Modelling years of service (YCS) as the dependent variable compresses the reference frame leading to better estimates with employees usually quitting within the first 10 years of service as shown in Figure 3.14b. Since employees will not quit if they are eligible for pension, those employees are removed from the risk set when they meet either of the requirements for retirement. With those cases now censored, the indicators P85 and A65 are no longer useful for modeling. Among all the models shown in Table 3.10, model



(a) **(b)**

Figure 3.14: Histogram of Age and YCS at Quitting

3 with DIV, OC, AGEC, and ERIP as explanatory variables fits best. A discrete version of the Cox model based on logistic regression also performs well with both models showing similar MAPE and G_2 values; see Allison (2010) for more details on discrete models.

The results suggest that voluntary quitting is influenced by an employee's age at the beginning of their service. The coefficient estimate for age is -0.025. Because the reference

Table 3.10: Model Assessment for Voluntary Quitting

No.	Dependent Variables	Independent Variables	LR	AIC	SBC	Pred. MAPE	Holdout MAPE	Pred. G^2	Holdout G^2
1	AGE	DIV GENDER OC YCSH ERIP	1226.8	5960.1	6041.4	83.70	91.35	1213.96	197.21
2	YCS	DIV OC AGEC ERIP	1012.4	6848.5	6929.7	23.90	21.05	65.63	2.48
3	YCS reduced riskset	DIV OC AGEC ERIP	838.4	6911.2	6992.5	15.16	18.77	26.25	2.08
4	Logistic regression	DIV OC AGEC ERIP	1870.7	4712.6	4938.6	15.98	17.55	23.03	3.21
5	Time series		NA	NA	NA	26.41	61.32	30.88	18.33

age is 35.44, the hazard of quitting for an employee who started working at age 36.44 is $exp(-.025) = .975$ that of an employee starting at age 35.44. For each additional year of age at the beginning of service, the hazard of quitting drops 2.5% . The employee's survival probability at any time, t, can be computed as $S(t)^{1.025} = (S(t)^{e^{0.025}})$ if that employee's age is one year below average, where $S(t)$ is the baseline survival probability for a reference employee of average age at initial employment. Moving in the other direction, the employee's survival probability increases to $S(t)^{0.975} = (S(t)^{e^{-0.025}})$ for a one-year increase in the employee's starting age. Together, these findings imply that at any given voluntary quitting age, the employee who starts earlier is more likely to quit than an equivalent employee who starts working later.

The early retirement incentive program (ERIP) also shows a significant impact on an employee's quitting behavior. This indicator's coefficient is 1.111, leading to a hazard ratio of $e^{1.111} = 3.04$. This ratio indicates that, on average, an individual's hazard of quitting increased by almost three times during this period. It is unclear why an optional early retirement program influenced quitting, but it is possible that the program led to leadership or organizational disruptions or that it simply occurred during 2008, a time of significant economic upheaval because of rapidly changing external economic conditions.

The DIV variable is also a significant predictor. For analysis, the baseline level was chosen arbitrarily as division 6 so that its hazard rate was determined by the baseline. Relative to this baseline, division 7 has a similar hazard of quitting according to the model estimates. Conversely, the other divisions have negative coefficients with hazard ratios less than 1,

Table 3.11: Parameter Estimates for Voluntary Quitting Models

		Model w/o external variable		Model with Real Dividend	
Parameter	Label	Parameter Est. (Std. Error)	Hazard Ratio	Parameter Est. (Std. Error)	Hazard Ratio
DIV	Div1	-3.066 (0.263)***	0.047	-3.305 (0.269)***	0.037
DIV	Div2	-2.652 (0.198)***	0.071	-2.875 (0.204)***	0.056
DIV	Div3	-3.015 (0.253)***	0.049	-3.258 (0.258)***	0.038
DIV	Div4	-2.533 (0.288)***	0.079	-2.769 (0.292)***	0.063
DIV	Div5	-2.739 (0.314)***	0.065	-2.934 (0.317)***	0.053
DIV	Div7	0.028 (0.145)	1.029	0.101 (0.146)	1.107
DIV	Div8	-0.806 (0.157)***	0.447	-0.968 (0.163)***	0.380
DIV	Div9	-3.985 (0.586)***	0.019	-4.207 (0.588)***	0.015
DIV	Div10	-1.325 (0.136)***	0.266	-1.500 (0.142)***	0.223
OC	C	-1.163 (0.274)***	0.313	-1.110 (0.275)***	0.329
OC	G	-0.794 (0.198)***	0.452	-0.796 (0.198)***	0.451
OC	L	-0.711 (0.228)**	0.491	-0.619 (0.229)**	0.539
OC	M	-0.541 (0.150)***	0.582	-0.497 (0.150)***	0.609
OC	P	-0.530 (0.135)***	0.588	-0.506 (0.136)***	0.603
OC	R	-0.950 (0.308)**	0.387	-0.886 (0.309)**	0.412
OC	S	-0.691 (0.275)*	0.501	-0.682 (0.276)*	0.506
OC	T	-0.608 (0.203)**	0.544	-0.564 (0.204)**	0.569
AGEC		-0.025 (0.005)***	0.975	-0.026 (0.005)***	0.974
ERIP	1	0.851 (0.135)***	2.343	0.479 (0.149)**	1.614
Real Dividends				0.085 (0.017)***	1.089

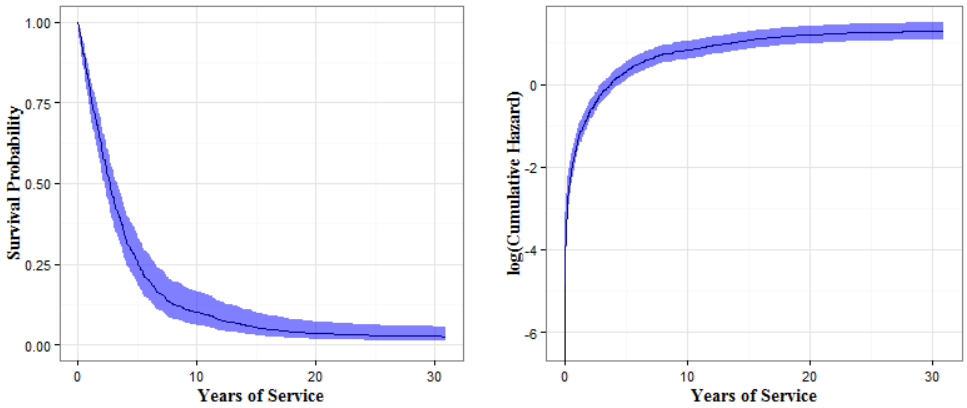
[1] * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$.

indicating that individuals within these groups have a lower hazard of quitting than group 6.

The OC variable is another significant predictor. For analysis, the baseline level is engineered for this variable so that its hazard rate is determined by the baseline. Relative to this group, the other divisions have negative coefficients with hazard ratios less than 1. In particular, the crafts group with a coefficient of -1.165 and a hazard ratio of 0.312 shows a much lower hazard of quitting than the engineer group. This finding may reflect differences in sociological or demographic factors among workers in these job categories, differences in compensation relative to other opportunities, or other differences in local or national economic mobility.

The baseline survival function and log cumulative hazard function for quitting are shown in Figure 3.15. The survival probability decreases steeply between 0 and 10 years of service. At 10 years of service, the survival probability has decreased to close to 0.25, indicating that 75% of employees quit within the first 10 years of service. From 10 to 30 years of service, the

survival function's slope flattens to 0. Accompanying the survival function is the cumulative hazard ratio's log. Again, the cumulative hazard's steep rise between 0 and 10 year of service indicates the increased quitting activity during this period. After ten years the cumulative hazard levels off, indicating a drop in the hazard rate at these future points.



(a) Survival Function **(b)** Log of Cumulative Hazard

Figure 3.15: Voluntary Quitting Model's Baselines with 95% Confidence Intervals

Predictive results for voluntary quitting without external variables appear in Table 3.7 columns 6 and 7. The quitting volumes are much lower than retirements in general, but average predictions are highly accurate for both training and holdout with the most sizeable deviation being general administration's under prediction by 2.6 units on average.

3.6.5 Voluntary Quitting Models including External Economic Variables

Because voluntary quitting is more sensitively affected by macroeconomics, the external variables are further examined using the counting process model with yearly interval based on the calendar year to test their effects on voluntary quitting. This model has the similar parameter estimation as the selected model as shown on the right side of Table 3.11.

The Real Dividend index is statistically significant and also improves the model forecasting because of a lower MAPE and G^2 than the selected model's values without economic indicator as shown in Table 3.12. The test results show that Real Dividend has

Table 3.12: Economic Index Test Statistics for Voluntary Quitting

Economic Inidcator[1]	χ^2	**P-value**	**Hazard ratio**	**MAPE**	G^2
Without Econmic indicator				15.71	28.33
MHP	37.93	<.001	1.012	13.62	18.44
SAMHP	37.75	<.001	1.012	13.64	18.50
SEMHP	39.80	<.001	1.020	13.74	18.67
SESAMHP	39.63	<.001	1.020	13.75	18.68
S&P500	0.02	0.879	1.000	15.46	27.79
Dividend	31.21	<.001	1.077	13.01	18.01
Earnings	8.83	0.003	1.009	15.81	23.88
Consumer Price Index	35.84	<.001	1.024	24.26	42.95
Real Price	5.34	0.021	1.000	17.94	34.52
Real Dividend	26.66	<.001	1.089	12.40	16.37
Real Earnings	3.71	0.054	1.005	14.16	23.34
Long Interest Rate	12.04	0.001	0.789	19.10	38.73
Unemployment Rate	2.99	0.084	1.068	17.04	34.96
P/E 10	16.22	<.001	0.968	17.95	35.19
Wilshire5000	10.75	0.001	1.031	15.84	21.31

[1] it is the selected model without economic indicator.

a strong impact on quitting behaviors. As shown in Figure 3.16, the voluntary quitting's plotted fluctuation corresponds with the trend of Real Dividend's one-year lag.

Predictive results for voluntary quitting with external variables appear in Table 3.7 columns 8 and 9. The predictions closely match those of the voluntary quitting model, which did not include external variables discussed in Section 3.6.4, thus indicating that while the model seems highly effective, the external variables' presence does not seem to substantially improve the predictions.

Table 3.13 shows the voluntary quitting model's yearly predictions along with bootstrapped confidence intervals. The third column shows the model predictions based on the fit using the Cox PH model. The bootstrap bias estimate indicates the difference between the predictions and actual mean values. Using ten thousands bootstrap replications, the bootstrap standard error is estimated. As shown in Table 3.13, the standard errors range from 2.1 to 7.8 for the voluntary quitting model. Although some actual quitting numbers are not included in the 95% prediction confidence intervals, they are close to the predicted confidence intervals. Also, the bootstrap bias estimates are all less than 1, indicating the predicted results generated by the Cox PH model are almost unbiased. The confidence intervals for voluntary quitting forecasting are shown in Figure 3.17.

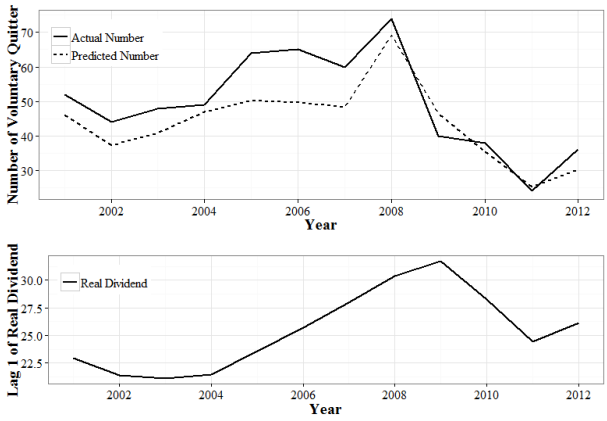
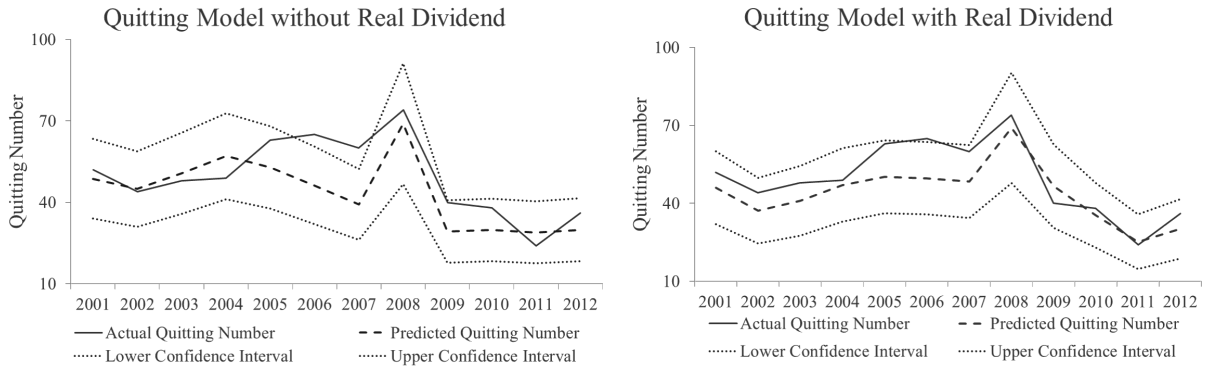


Figure 3.16: Actual vs. Forecast by Voluntary Quitting Model with Real Dividend and Its Index Plot

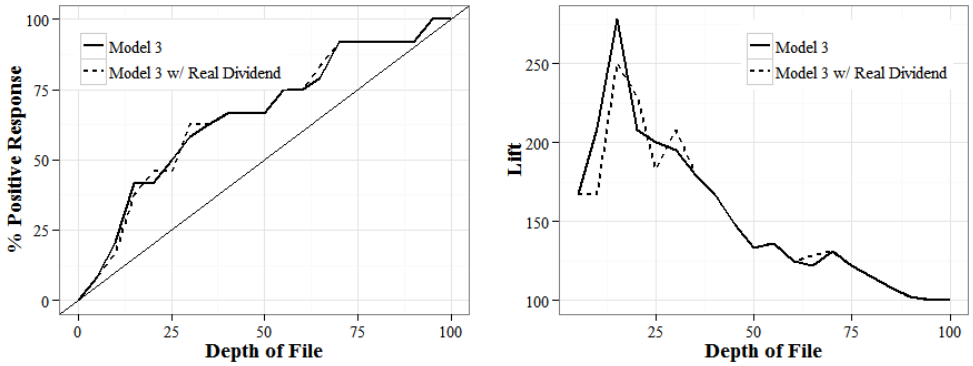
Table 3.13: Voluntary Quitting Models' Bootstrapping Results

		VQ w/o Real Dividend				VQ w/ Real Dividend			
Year	Actual	Pred.	Bias	Std. Error	$Var\sum\hat{Y}_i$	Pred.	Bias	Std. Error	$Var\sum\hat{Y}_i$
2001	52	49	0.1948	2.79	56.49	46	-0.0098	2.71	51.05
2002	44	45	0.1256	2.30	50.26	37	-0.0364	2.47	41.35
2003	48	51	0.2028	2.66	57.83	41	-0.0549	2.89	46.69
2004	49	57	0.3166	2.80	64.83	47	-0.005	3.06	52.19
2005	63	53	0.1689	2.63	59.70	50	-0.0889	2.58	51.03
2006	65	46	0.2658	2.58	52.84	50	-0.0815	2.54	50.41
2007	60	39	0.1600	2.28	44.48	48	0.0376	2.84	51.23
2008	74	69	-0.3258	7.78	129.59	69	0.2345	7.35	117.52
2009	40	29	0.1049	2.23	34.24	47	0.0522	4.70	66.56
2010	38	30	0.0827	2.14	34.35	35	-0.0600	2.42	40.04
2011	24	29	0.1383	2.25	34.01	25	-0.1132	2.09	28.96
2012	36	30	0.1514	2.26	35.05	30	-0.0774	2.15	33.88



(a) **(b)**

Figure 3.17: Voluntary Quitting Models' Forecast with 95% Confidence Intervals



(a) Cumulative Gain Chart **(b)** Cumulative Lift Chart

Figure 3.18: Voluntary Quitting Models: Gain Chart and Lift Chart for 2011 Holdout

As with the retirement models, lift and gains plots can also be used to assess the predictive models effectiveness for quitting by using out-of-sample predictive accuracy. Figure 3.18 compares the optimal voluntary quitting model (model 3 from Table 3.10) against the same model with the external variable Real Dividends added. Gains and lift charts are interpreted as they were in the retirement case discussed in Section 3.6.3. The gains figure indicates that both models with and without Real Dividends perform with similar predictive accuracy at the individual level, indicating that including the external variable has little practical impact on the model. The number of employees quitting during the sample window is lower than observed for retirement; and, given the information provided by the data, predicting quitting at the individual level is more difficult.

After sorting predicted probabilities, among the 25% of employees predicted most likely to quit in the 2011 holdout sample, 50% of the actual quitting cases were observed. Furthermore, the top 50% of predictions contained about 66.7% of the quitting cases, and the top 60% contained close to 75% of the actual cases. Results indicate that the model including Real Dividends predicts well but is not as effective as the model without those dividends. In either case, the model does provide predictions of quitting propensity that can be useful.

Figure 3.18b provides lift estimates, indicating that among the files top 25%, model 3 captures twice as many of the actual quitting cases as a random sample of the same size. The model including Real Dividends performs similarly. Based on this finding, the model provides

a useful target group for human resource administrators, allowing them to modify and focus their retention strategy for maximum effect. It also identifies an enriched population for management researchers to study the reasons why workers decide to quit.

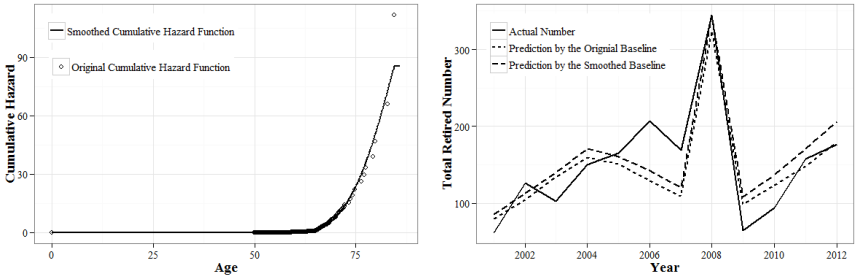
3.6.6 Baseline Smoothing Results

3.6.6.1 Retirement Baseline Smoothing Results

The selected retirement model without the financial indices baseline was smoothed. As shown in Figure 3.19a, a smoothed cumulative hazard function was generated across the original cumulative hazard. The smoothed values are close to the actual when age is from 50 to around 70 years old because most employees retired before they were 70 years old. After age 70, the smoothed baseline slightly over estimates the actual value, and it greatly underestimates the cumulative hazard when an employee retires at age 84. The predicted results showed the smoothed baseline always predicts a larger number than the original baseline as shown in Figure 3.19b. The prediction has a smaller gap between the smoothed baseline and the actual retiring number at 2008. However, the prediction has a larger deviation using the smoothed baseline than the one without it in the other years. Furthermore, the training and holdout G^2 are 103.1 and 5.7, and overall G^2 is 108.8. Also, the training and holdout MAPE are 27.3% and 12.5%. Recall G^2 in Table 3.5, the selected model's G^2 values are 111.6 and 0.8 for training and holdout, respectively. The smoothed model performed better in training dataset, but over-predict in the holdout sample. Therefore, the baseline smoothing methods did not provide a better prediction for retirement model.

3.6.6.2 Voluntary Quitting Baseline Smoothing Results

The cumulative function was smoothed from the selected voluntary quitting model. In Figure 3.20a, the solid line is the smoothed cumulative function, and the dot is the original voluntary quitting model. The smoothed values are lower than the original values in the first year. Also. the smoothed function underestimates from 7 to 10 years of service. Other than those years, the smoothed function is close to the original one. Figure 3.20b compares the prediction using

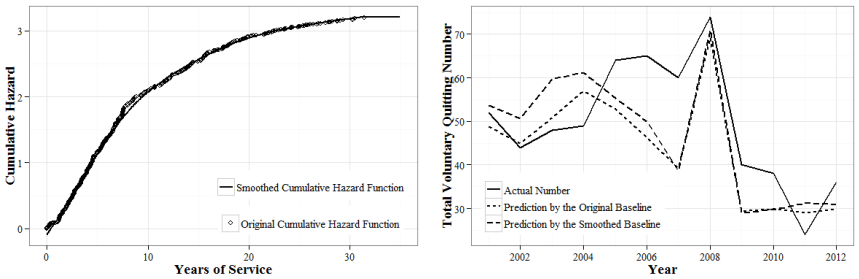


(a) Smoothed Cumulative Hazard Function

(b) Retirement Forecast Comparison: Original vs. Smoothed Baselines

Figure 3.19: Retirement Model's Smoothed Baseline Plot and Forecast Comparison by Original and Smoothed Baselines

the smoothed baseline and the original baseline. The smoothed baselines prediction always surpassed the original baseline. Only in 2008, the smoothed baselines estimation is closer to the actual value. In the other years, this estimation deviates more from the actual values. Furthermore, the training and holdout G^2 are 27.8 and 2.7, respectively, and the overall G^2 is 30.4. Also, the training and holdout MAPE are 17.6% and 22.2%. Recall G^2 in Table 3.10, the selected model's G^2 values are 26.36 and 2.8 for training and holdout, respectively. The smoothed model did not perform better in either training or holdout dataset. Therefore, the baseline smoothing method attempts to overestimate the actual turnover values in both the retirement and voluntary quitting models in this study.



(a) Smoothed Cumulative Hazard Function

(b) Voluntary Quitting Forecast Comparison: Original vs. Smoothed Baselines

Figure 3.20: Voluntary Quitting Model Smoothed Baseline Plot and Forecast Comparison by Original and Smoothed Baselines

3.7 Conclusions and Managerial Implications

Using the Cox proportional hazards model along with appropriately chosen internal and external variables led to accurate retirement predictions. In the training sample, prediction error as measured by MAPE was approximately 25% while the predictions in the smaller two-year holdout window were approximately 5%. Although including only two validation points, these results indicate that the method has favorable potential.

The key internal variables that improved model predictions include division, years of service at hire (YCSH), and age at hire (AGEH). In addition, S&P 500 real earnings were also significantly associated with risk of retirement although the magnitude at approximately 1.3% was (greater or less) than for small changes' other effects. The retirement hazard increases significantly when an individual accumulates 85 points of retirement credit and is eligible for full benefits but then reverts to a lower hazard after age 65. Furthermore, the early retirement incentive plan that the organization implemented in 2008 had a major impact with a large increase in the hazard of retirement, particularly for those eligible for full benefits with points exceeding 85. Considering individual and external information in out-of-sample tests provides a significant improvement over more traditional forecasting methods Feldman (1994). Such information also provides useful predictions on subgroups that are not possible with standard forecasting models.

Quitting behavior differed significantly by division, occupation, and age at start of service, reflecting differences in both worker satisfaction across the organization and by job type. Age connects logically because for a given number of years of service, the employee who starts earlier is younger and, therefore, may see a more significant long-term opportunity in a new position elsewhere. The 2008 early retirement incentive program was also correlated with a significant amount of quitting. It is unclear if the higher hazard of quitting resulted directly from disruptions, management, accelerated retirements, or the drastic economic changes during that time. In terms of predictive power, the quitting model performs well with MAPE in the training sample estimated at 15.16% and at 18.77% in the holdout sample. Both estimates are far superior to those that traditional forecasting techniques generate. As with retirement, external information in the form of S&P real corporate dividends were

positively correlated with quitting, thus indicating that as large private-sector companies' profits increase, the hazard of quitting rises significantly.

Drawing on these findings, this work provides a number of valuable managerial implications. Foremost, this work shows that fairly accurate survival model-based retirement prediction is feasible in large organizations when defined benefit plans are in place. These models are accurate enough to provide useful predictive visibility for human resources management professionals. Such models are of particular value when a large portion of the workforce has specialized skills requireing an extensive search process to replace or if the organization requires long lead times in the hiring process because of security or other concerns. The model also provides guidance as to expected retirement and quitting behavior across the organization's subgroups. Significant deviations from these expectations, particularly in terms of abnormal quitting behavior, may indicate the need for management intervention. Finally, such models can alert managers to key external factors that may indicate increased retirement or quitting behavior. Using lagged variables may allow management to craft incentive policies in response to changes in the external economy.

Although this study brings to light much, it is important to consider a number of limitations because of the data and models used. The data were sampled based on a specific time window leading to a truncated sample, which limits the modeling approaches. Thus, it would be beneficial to have a longer prediction window to validate the model. Variables such as salary and more carefully documented division data, which would be available within the sponsoring organization, could strengthen the results presented here. In the same vein, more complete internal organizational details on early retirement incentive program as well as the defined benefit plan's details could improve the inferences' quality in this study. Finally, the accuracy of results for this study's retirement portion are based on the specific nature of the defined benefit pension dominant during the study period. The accuracy of predictions observed in this study may not generalize to organizations offering defined contribution retirement plans although a similar methodology may still prove useful in generating predictions. Although tested, the proportional hazards assumption may be violated for some effects in terms of the model. In addition, the baseline estimator's

variability can cause overestimation of hazard. Thus, more sophisticated smoothed baseline models may be beneficial in this case.

Using predictive models, this work contributes significantly to both practical and academic literature regarding factors affecting retirement and management of staffing. Future work should continue developing and improving the modeling techniques and result in better explanatory models for academic research as well as better functioning systems for industry. Furthermore, it is believed that the current methods can be extended to include qualitative and survey-based longitudinal feedback from employees through the time-varying covariate methodology. The addition of such attitudinal data may provide significant value to both academic and industry practitioners.

Chapter 4

Employee Voluntary Turnover Determinants Analysis and Forecasting for R&D Departments

4.1 Introduction

Researchers and human resource managers have focused on employee turnover for decades because it negatively affects organizations' performance (Shaw, 2011). Employee retention is one of the main challenges in organizations, especially for those with a long lead time to hire a new employee. In this study, besides the hiring and training time, the security background check takes a long time ($>$ 6 months) when filling positions. Our study's subjects were 731 terminated employees in four research and development departments across more than a ten-year study window. This study focuses on employee voluntary turnover, i.e., employees who voluntarily quit their job. Among the reasons for termination, voluntary turnover is one of the major ones, accounting for 26%. Compared to retirement and layoff, voluntary turnover is harder for companies to control. Because voluntary turnover is expensive for companies (Selden and Moynihan, 2000), they do not want their employees to voluntarily leave (Allen et al., 2010). Therefore, this study's objective is to examine what determines

voluntary turnover for R&D department employees and to forecast employee tenure, thus assisting in hiring decisions.

4.2 Literature Review

Employee turnover represents a major loss of intellectual property to an organization. The total cost associated with employee turnover is estimated to range from 100-300% of a departing employee's annual salary (Moody, 2000). Both private firms and governmental organizations spend billions of dollars every year to manage employee turnover (Leonard, 2001). For organizations to execute their tasks efficiently and effectively with the highest quality, they must ensure that the right people are available at the right places and at the right times (Khoong, 1996). Many researchers have attempted to identify turnover factors to prevent and reduce turnover. Bluedorn (1982) reported that turnover is related to an individual's, routine, age, length of service, and perception of environmental opportunities. Balfour and Neff (1993) suggested that caseworkers with more education, less experience, and less at stake in an organization are more likely to turnover. According to Griffeth et al. (2000), pay and pay-related variables modestly affect turnover. They also examined the relationship among pay, a person's performance, and turnover. They concluded that when high performers are inadequately rewarded, they quit. In our dataset, company job titles are highly and positively correlated with employees' salaries, sensitive information that cannot be released. We can hypothesize the following:

Hypothesis 1: The company job title is a significant predictor of voluntary turnover.

Besides job titles, we think employees' college majors are also a factor causing variation of employees' salaries. Thus, we hypothesize the following:

Hypothesis 2: The college major (branch of study) is a significant predictor of voluntary turnover.

Griffeth et al. (2000) study shows that employees' demographic characteristics such as gender and education level are also highly correlated to employee turnover. They also found that men are less likely to turnover than women. Thus, we hypothesize the following:

Hypothesis 3: Employee education level a significant predictor of employee voluntary turnover.

Hypothesis 4: Female employees are more likely to voluntary turnover than male employees.

Many researchers have examined the relationship between employees' age and turnover. Younger employees are more likely to move from one job to another (Burke, 1994), particularly those less than 35 years old (McShane and Von Glinow, 2003). Thus, we hypothesize the following:

Hypothesis 5: Age is a significant factor of employee voluntary turnover.

Employees are more likely to quit during the first five years before they have a strong commitment to the organization. Bluedorn (1982) found that turnover is related to length of service. Thus, we hypothesize the following:

Hypothesis 6: Length of service is a significant factor of employee voluntary turnover.

Furthermore, some races may react differently to other races. For example, Thaden et al. (2010) found that employee turnover is significantly related to race. Ethnic minorities' language and accent barriers may cause social differences with the ethnic majority. Thus, we hypothesize the following:

Hypothesis 7: Ethnicity is a significant predictor of employee voluntary turnover.

Many researchers have used logistic regression to build models and to identify turnover's attributes. For example, Balfour and Neff (1993) applied logistic regression to build an employee turnover model. Wright and Cropanzano (1998) used correlation and logistic regression to examine whether emotional exhaustion is a predictor of turnover. Morrow et al. (1999) applied logistic regression to determine whether employees' absence and performance record can be employee-turnover indicators. Nagadevara et al. (2008) applied several statistical methods including logistic regression to predict employee turnover. To predict employee turnover behavior, many other statistical techniques (such as regression, neural network (NN), and data mining) have also been used. For example, Ng et al. (1991) used a proportional hazards regression (PHR) to develop a turnover prediction model.

Beng Ang et al. (1994) developed a turnover prediction model for accountants in a Singapore organization. Jenkins and Paul Thomlinson (1992) used multiple regression to explore turnover intention. Chang and Xi (2009) combined Taguchi's method and Nearest Neighbor Classification Rules to select feature subsets and analyze factors to find the best predictor of employer turnover. Alao and Adeyemo (2013) used a decision tree to classify employees based on various types of turnover using employees' records. However, all these modeling methods have attempted to forecast employees' turnover behaviors rather than providing a forecast model or rules that forecast employee tenure and that also assist in hiring employees who are less likely to voluntarily quit.

This chapter has five parts. Part one is the introduction including the objective and literature review. Part two provides a synopsis of tools for data preparation and discusses forecasting methods. Part three presents results. Part four discusses the results. Part five is the conclusion.

4.3 Methodology

4.3.1 Data and Preparation

In 2011 a large U.S. organization provided the human resource data for this study. The dataset contained 731 observations of terminated employee records from October 2000 to June 2011, including metrics such as ID, division (department), job title, termination reason, last hired date, termination date, years of service (YCS), gender, race, age at hiring, age at termination, highest education degree, and branch of study. In addition, two variables were created from the dataset for analysis. One variable was major, including 15 levels. The first three letters of the branch of study are used for identification; for example, Eng refers to engineering, and Che refers to chemistry. The other was a binary target variable Y for logistic regression. Y is equal to 1 if an employee voluntarily quits; otherwise, y is equal to 0 if an employee leaves the organization for other reasons.

4.3.2 Logistic Regression

The hypotheses were tested using logistic regression to identify voluntary turnover's significant predictors. Logistic regression, or logit regression, is a popular model for binary data (Agresti, 1996). It is used to predict a binary response from a binary predictor, to predict a categorical variable's outcome based on one or more predictor variables, and to estimate a qualitative response model's parameters. Unlike linear regression, logistic regression can handle both categorical and numeric variables. The logistic regression is expressed in Equation 4.1

$$logit[\pi(x)] = log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x \tag{4.1}$$

where x represents the explanatory variables; $\pi(x)$ denotes the target variable's probability at value x; $\frac{\pi(x)}{1-\pi(x)}$ is the odds of x; and α, β are the estimated coefficients. All the predictors were tested using 0.05 as a p-value criterion. The other variables, such as department, were also included in the model to eliminate the other variables' variance. Three statistics criteria were used to evaluate the logistic regression model: Akaike's information (AIC), Schwartz's Bayesian criterion (SBC), and the Hosmer-Lemeshow goodness of fit test. AIC and SBC are both information criteria using likelihood value. Usually, the best model comes with the lowest AIC or SBC values. The Hosmer-Lemeshow test is a chi-square test to determine whether the logistic regression model is correctly specified (Hosmer Jr and Lemeshow, 2004). The high p-value ($P > 0.05$) indicates that this model passes the test.

4.3.3 Decision Tree

A decision tree, also called a classification tree, is one method of classification via an intermediate tree-like structure in data mining (Hand et al., 2001). A decision tree's purpose is to develop a classification rule from the data based on attributes, or explanatory variables. Similar to logistic regression, a decision tree can handle categorical or numeric variables. Creating a decision tree involves splitting the significant variable until only one unique classification is on each branch of the tree. The decision tree was used for data compression and prediction. In this study, YCS were used as a target variable generating a classification

rule for managers and human resource departments to forecast how many years an employee will work in the organization. It also provides managers a tool for making accurate hiring decisions. The turnover dataset was divided into two parts: training (60%) used to build a decision tree and validation (40%) used to validate the model performance. The decision tree model was built using AIC values as model selection criteria. The decision tree was analyzed using SAS Enterprise Miner 7.

4.4 Results

4.4.1 Hypothesis Test Results

All variables proposed in the hypothesis were tested with other control variables using the logistic regression model. The parameter estimations are shown in Table 4.1. The variables included Div (division), jobtitle (job title), race, gender, degree (highest education degree), major (branch of study), age (age at termination), and YCS (years of company service). Five variables were statistically significant ($P < 0.05$) in the logistic regression model: jobtitle, gender, race, age, and YCS. This finding indicates that employees voluntarily quitting the organization are statistically affected by their company job title, gender, race, age, and years of company service.

Table 4.1: Logistic Regression Parameter Estimates

Effect	DF	Waldχ^2	$Pr > \chi^2$
Jobtitle	14	37.21	<.01
Division	3	4.07	0.25
Gender	1	10.72	.001
Degree	5	2.25	0.81
Major	14	9.75	0.78
Race	4	30.56	<.01
Age	1	38.81	<.01
Years of Service	1	45.62	<.01

The results of the hypotheses tests are discussed below.

Hypothesis 1: The company job title is a significant predictor of voluntary turnover.

Table 4.2: Logistic Regression Parameter Estimates 2

Parameter	Levels	DF	Estimate	Standard Error	Wald χ^2	Pr $>\chi^2$
Intercept		1	10.605	2.039	27.048	<.0001
Job title	Administrator1	1	-5.710	1.589	12.912	0.001
Job title	Administrator2	1	-2.157	1.166	3.425	0.064
Job title	Administrator3	1	-2.892	1.155	6.265	0.012
Job title	Administrator4	1	-0.042	1.089	0.002	0.969
Job title	Staff 1	1	-1.788	0.966	3.428	0.064
Job title	Staff 2	1	-0.688	0.803	0.733	0.392
Job title	Staff 3	1	0.065	0.667	0.009	0.923
Job title	Staff 4	1	0.877	0.688	1.622	0.203
Job title	Staff Member1	1	-1.479	1.216	1.479	0.224
Job title	Staff Member2	1	0.935	0.983	0.903	0.342
Job title	Staff Member3	1	-0.683	0.793	0.741	0.389
Job title	Staff Member4	1	0.907	0.862	1.108	0.293
Job title	Technician	1	-2.652	0.992	7.146	0.008
Job title	Others	1	-11.141	820.7	0.000	0.989
Gender	F	1	1.293	0.395	10.720	0.001
Race	Asian	1	-1.489	0.632	5.545	0.019
Race	Black or African American	1	-3.917	0.733	28.595	<.0001
Race	Hispanic/Latino	1	-1.495	1.068	1.959	0.162
Race	Native American/Alaskan	1	12.905	1679.0	0.000	0.994
Age		1	-0.129	0.021	38.812	<.0001
YCS		1	-0.149	0.022	45.618	<.0001

The company job title is a pivotal factor in voluntary turnover since it is highly associated with employee salary ($\chi^2 = 37.2, df = 14, P < 0.05$). We have no evidence to reject the null hypothesis. Compared to the reference group (Managers), four job title groups are statistically significantly different than the manager group. These company's job titles are Administrator 1, Administrator 2, Administrator 3, Staff Level 1, and Technician; their coefficient estimates are -6.7, -2.9, -3.6, -1.8, and -3.0, respectively. These results show that compared with managers, employees in these positions are less likely to quit: Administrator 1 has 99.9% (1-exp(-6.7)) less odds than the reference group; Administrator 2 has 94.5% less odds than the reference group; Administrator 3 has 97.3% less odds than the reference group; Staff 1 has 83.4% less odds than the reference group; and Technician has 95% less odds than the reference group as shown in Table 4.2. The other job titles are not statistically significant, indicating they have the same turnover odds as managers.

Hypothesis 2: The college major (branch of study) is a significant predictor of voluntary turnover.

According to Table 1, employees' major has 9.7 of Wald χ^2 with 14 degrees of freedom, and P value is 0.78. We have evidence to reject the null hypothesis. Employees' college

major (branch of study) does not significantly affect employee turnover behaviors in R&D departments.

Hypothesis 3: Employee education level is a significant predictor of employee voluntary turnover.

In this study, an education degree is not a significant predictor of employee voluntary turnover ($\chi^2 = 2.25, df = 5, P = 0.81$). We have evidence to reject the null hypothesis, thus showing that employees with a higher education degree have voluntary turnover behavior similar to those without a higher education degree.

Hypothesis 4: Female employees are more likely to voluntary turnover than male employees.

We do not have evidence to reject the null hypothesis. Gender is also significant ($\chi^2 = 10.7, df = 1, P < 0.05$), indicating that female and male employees have different voluntary quit behaviors, as the parameter estimation for a female employee is 1.29. Also, the odds of female voluntary turnover are 3.6 times that of male employees, indicating female employees are more likely to voluntarily quit.

Hypothesis 5: Age is a significant factor of employee voluntary turnover.

We do not have evidence to reject this null hypothesis. Employees' age significantly affects voluntary turnover ($\chi^2 = 38.8, df = 1, P < 0.05$). The coefficient estimate is -0.13 and the odds ratio is 0.87, indicating that the turnover odds decrease 13% when an employee's age increases one year older. As a result, younger employees tend to be more likely to quit than older employees.

Hypothesis 6: Length of service is a significant factor of employee voluntary turnover.

We do not have evidence to reject this null hypothesis. Employees' years of service significantly affect voluntary turnover ($\chi^2 = 45.6, df = 1, P < 0.05$). The coefficient estimate is -0.15 and the odds ratio is 0.86, indicating that the turnover odds decrease 14% with each additional year of service. Thus, employees with fewer years of service tend to be more likely to quit.

Hypothesis 7: Ethnicity is a significant predictor of employee voluntary turnover.

We do not have evidence to reject this null hypothesis. Ethnicity has a statistical impact on employee voluntary turnover behavior ($\chi^2 = 30.6, df = 4, P < 0.05$). Asian, Black or African American, and Hispanic/Latino are significantly different statistically from white employees, the reference group, in terms of turnover because these three groups have negative coefficient estimates. Black or African American employees have the lowest coefficient estimates (-3.7) and odds ratio (0.024=exp(-3.7)). Asian and Hispanic/Latino employees have -1.3 and -1.5 of coefficients and 0.27 and 0.22 of the odds ratio, respectively. These three groups have lower turnover probability than white employees. However, Native American employees are not significantly different statistically from white employees ($P > 0.05$), showing that Native American employees have the same turnover probability as white employees.

4.4.2 Decision Tree Analysis Results

This study's decision tree model is a five-level depth tree with 17 nodes and has the lowest AIC value as shown in Figure 4.1. Four variables are statistically significant in the model: *ageh* (age at hire), *jobtitle* (job title), *div* (division), and *race*. The most important variable is *ageh* located in the decision tree's first, third, and fourth nodes, indicating that an employee's YCS is mainly determined by age when hired. Two other variables, job title and division, also significantly affect an employee's YCS. Based on the model, DIV1 is significantly different from the other three departments (DIV2, DIV3, and DIV4). The employees with the following job titles are significantly different from those with other job titles: Staff 2, Staff Member 1, Administrator 2, or Administrator 1. Compared to other ethnicities, white employees have more YCS if they are younger than 28 when hired. The decision tree's rules are the following:

- if $ageh \geq 42.25$ and div =DIV1, then $E(YCS) = 11.9$;
- if $ageh \geq 42.25$ and $div \neq$ ES1, then $E(YCS) = 5.6$;
- if $21.35 \leq ageh < 42.25$, $jobtitle$ = (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), and div =DIV1, then $E(YCS) = 22.9$;

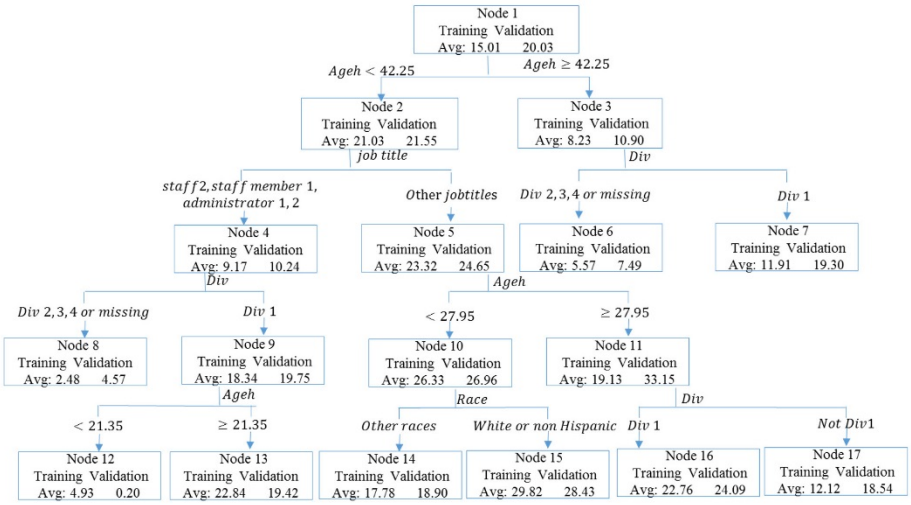


Figure 4.1: R&D Employee YCS Decision Tree Model

- if $ageh < 42.25$, $jobtitle =$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), $div =$DIV1, and $ageh < 21.35$, then $E(YCS) = 4.9$;
- if $ageh < 42.25$, $jobtitle =$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), and $div \neq$ DIV1, then $E(YCS) = 2.5$;
- if $27.95 \leq ageh < 42.25$, $jobtitle \neq$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), and $div =$ DIV2, DIV3, or DIV4 then $E(YCS) = 12.1$;
- if $27.95 \leq ageh < 42.25$, $jobtitle \neq$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), and $divneq$ DIV2, DIV3, or DIV4 then $E(YCS) = 22.8$;
- if $ageh < 42.25$, $jobtitle \neq$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), $ageh < 27.95$, and $race =$ white or white/non-Hispanic origin, then $E(YCS) = 29.8$;
- if $ageh < 42.25$, $jobtitle \neq$ (Staff 2, Staff Member 1, Administrator 2, or Administrator 1), $ageh < 27.95$, and $race \neq$ white or white/non-Hispanic origin, then $E(YCS) = 17.8$;

where E(YCS) denotes expected or average YCS. The predicted YCS's range is from 2.5 to 29.8 years. The model's performance is validated by predicting E(YCS) in the validation

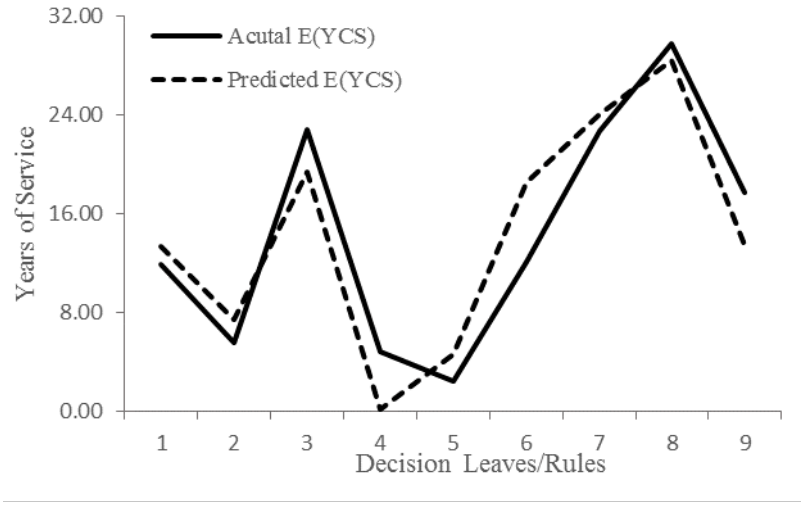


Figure 4.2: Actual vs. Predicted E(YCS) for Validation Dataset

dataset. As shown in Figure 4.2, the predicted values are close to the actual values in each leaf, indicating the decision tree model has strong forecasting ability. After the decision rule is determined, predicting YCS for a new employee is easily calculated by following the rules. This ease of calculation is one of the decision tree method's advantages.

4.5 Discussion

This study's purposes were to investigate factors contributing to R&D employee voluntary turnover and to support employee hiring and retention strategies. The logistic regression's results show that job title, gender, ethnicity, age, and years of service are significant predictors of employee turnover. However, an employee's division, education level, and major are not associated with employee turnover.

Although we found that employee job title is a significant predictor of employee turnover, the comparison results show that employees with lower salaries are more likely to stay. This finding is different from the previous study. Compared to the other staff levels, level 1 staff are also more likely to stay because they do not have much work experience. Also, their average salary is 20% higher than the job market's average. On the other hand, the other staff levels have the same probability to quit as the reference group (managers). According to Tuji (2013) study, when employees find no opportunities for advancement within the system,

they will not remain in the work situation. The promotion competition is intense among R&D staff members with the same education degree and experience. However, they easily find higher paying positions in other companies or become members of faculty or research staff in organizations when they have several years of work experience. Therefore, R&D staff are more likely to quit when they are not in staff level 1.

Furthermore, female employees are more likely to quit than male employees. According to the work/family life-balance theory (Clark, 2000), as economic pressure increases, an employee may have to do the work of more than one person. Especially when an organization downsizes or restructures, the same amount of work has to be done by fewer employees. Thus, employees spend more time on their work life and have less time for their personal life (Smith, 2009). Female employees serve a more important role in the family than male employees. Forced to make a choice between family and work, female employees are more likely to quit.

Many studies have proved that age has a negative relationship with turnover (Rhodes, 1983). Researchers believe younger workers leave for two main reasons: lack of training opportunities and lack of mentors in the workplace (Paul, 2012). Although many managers are promoted because of their strong capability and achievement, they are not necessarily good coaches or team leaders and are unable to help employees improve their performance (Smith, 2009). Furthermore, R&D employees work more independently than employees in other departments. For example, sometimes one employee is responsible for one project or part of a project. Compared to experienced staff, younger staff will feel frustrated and stressed when they meet difficulties or barriers and cannot receive adequate advice. As a result, younger employees are more likely to voluntarily leave.

The number of years an employee has worked in the organization is another significant factor in turnover. Many studies have found that length of service is highly related to organization commitment (Jena, 2015; Kelarijani et al., 2014; Popoola, 2006). The longer the employee works, the more that employee feels attached to and involved in the organization. As a result, employees with longer length of service are more likely to stay. On the other hand, employees with few years of service have not formed a strong commitment to the organization. Thus, they tend to switch to a new job to increase their work experience, receive higher pay, or be in a higher position.

Ethnicity is another significant factor affecting employee turnover. According to our analysis, Black or African American employees have the lowest probabilities of quitting compared to the other ethnic groups. Asian and Hispanic employees are also less likely to quit than the white employees in the study. However, Native American/Alaskan employees have the same turnover probability as white employees.

Employee turnover does not associate with majors as shown in Table 1. Although some majors like computer science or engineering have more job openings, they do not have a higher turnover probability in R&D departments because employees have 20% higher salaries than the job market's average levels, according to our study.

Education degree is not a significant factor in voluntary turnover. Most employees (80%) in the four R&D departments are highly educated with at least a bachelor's degree. However, technicians and administrators are more likely to have a high school or an associate degree. As discussed above, they are all less likely to quit. Therefore, determining which employees with which degree are more likely to quit is difficult. Employee division in this study is not significant ($\chi^2 = 4.7, df = 3, P > 0.05$). The results indicate that employees have the same turnover probability among the four divisions, which have similar structures and cultures.

The decision tree model predicts the average employee's working years in the R&D department under all types of employee turnover. The significant variables−age when an employee was hired, job title, division, and ethnicity−are easily obtainable. Based on the decision rules, managers can quickly determine whether an interviewee will have a long or short tenure. Based on the rules shown in section 4.2, a new employee's years of service can be predicted. If a new employee is applied to rules (2), (4), and (5), that employee's predicted YCS will be fewer than 10 years. If a new employee is applied to rules (1), (6), and (9), the predicted YCS will be 10 to 20 years. If a new employee is applied to rules (3), (7), and (8), the predicted YCS will be above 20 years.

4.6 Conclusions and Recommendations

In this study, employee turnover's significant attributes and classification rules were identified based on employee records from four R&D departments. Seven variables in the hypothesis

were tested in the logistic regression model to find the variables' significance. This logistic regression determined the probability of an employee's voluntary turnover and identified four significant factors affecting voluntary turnover: job title, gender, ethnicity, age, and years of service. These results can assist managers and human resource departments in developing employee-retention strategies to reduce R&D departments' voluntary turnover rate.

The decision tree generated nine rules to predict the average length of YCS of an employee. The models' results showed that combining models were suitable for forecasting employee turnover. Applications of the models can be used with hiring strategies. For example, when the data and related variables are accessible, a decision tree can generate a decision rule, such as hiring or not hiring an applicant. Results from the decision tree indicated that the age of employees when hired was the most important variable.

Several statistical software packages are available to conduct logistic regression and decision tree models, such as SAS, SPSS. If human resource departments have limited budgets, they can also use statistical software R, which is a free open source. These two models are created based on the termination dataset with limited variables. If more data and variables can be identified, the models can be further improved.

Chapter 5

Conclusions

5.1 Summary of Findings

In this study, various time series models for forecasting employee-turnover counts were tested, and optimal models for turnover forecasts were identified. The dynamic regression model with the additive trend, seasonality, interventions, and U.S monthly composite leading indicator (CLI) efficiently predicted the turnover with training $R^2 = 0.77$ and holdout $R^2 = 0.59$.

The Cox proportional hazards model along with appropriately chosen internal and external variables led to accurate retirement prediction predictions. In the training sample, training MAPE is approximately 25% while the predictions in holdout data are approximately 5%. The key internal variables are division, years of service at hire (YCSH), and age at hire (AGEH). Retirement hazard increases significantly when an individual hit 85 points of retirement credit but then reverts to a lower hazard after age 65. Furthermore, the early-retirement incentive program the organization implemented in 2008 significantly increases the hazard of retirement, particularly for those with points exceeding 85. In addition, S&P 500 real earnings are significantly associated with hazard of retirement. Quitting behavior differs significantly by division, occupation, and age at start of service. 2008 early-retirement incentive program is also correlated with a significant amount of quitting. The quitting model performed well with MAPE in the training sample estimated at 15.16% and at 18.77% in the holdout sample. In addition, external information, S&P real

corporate dividends, is positively correlated with quitting, indicating that as large private-sector companies' profits increase, the hazard of quitting rises significantly.

Also, this study identified not only significant factors related to employee turnover but also classification rules based on employee records from four R&D departments. Four significant factors affecting voluntary turnover are job title, gender, ethnicity, age, and years of service. However, employees' major, education level, and department do not correlate with R&D employees' quitting behaviors. The decision tree generates nine rules to predict a new employee's tenure. These results can assist managers and human resource departments in developing employee-retention strategies, while also reducing hiring lead time and employee turnover cost.

5.2 Implementation of Employee Turnover Forecasting Models

The conceptual models for building an employee-turnover forecasting system includes three steps: model installation and forecast-results identification, strategy generation, and model update. These steps are discussed below.

The first step is installing the employee-forecasting model in the HR management system to forecast employee turnover. The forecasting model must be easy to install and have a user-friendly interface. After the model is installed, the quarterly or yearly employee-turnover records are exported from the HR management system. Those records include employees' individual information, such as gender, age, years of service, department, and job title. Then this information is imported into the model, and the predicted employee-turnover number (P) for the next six or twelve months is computed. Based on the production plan and budget, HR calculates the number of employees needed in a future year as the demand (D), identifies the current number of employees (S), and finally, HR can compute the number of employees required in the next year as D-(S-P). The employees predicted to have a high probability of turnover are targeted for HR to determine retention and promotion strategies.

The second step is modifying employee-retention strategies and building a talent-inventory control system. Moncarz et al. (2009) found that effective retention strategies, such as promotions and training, can reduce employee turnover long term and positively influence employee retention and tenure. The significant turnover factors provide managers a clear direction for creating retention strategies in such areas as corporate culture and communication, work environment and job design, promotions, customer contentedness and employee recognition, rewards, and compensation (Moncarz et al., 2009). However, some organizations with good retention plans still experience high employee turnover as a result of a high proportion of aging workforce. The optimal employee-inventory management strategy reduces cost and maintains a lean management system's normal operation. Many inventory models, such as the Economic Order Quantity model (EOQ), can be used to manage employee inventory. The EOQ model assumes the demand for the employee inventory occurs at a constant rate; an ordering and setup cost k is incurred when hiring new employees; the cost per-year of holding employee inventory is h; and no shortage of employees is allowed. Realistically, the employee-demand number can be determined by the current workforce number, the total turnover counts, and the future workforce number. The company's production plan and strategy determine the total workforce required, and the employee-turnover forecasting model provides the future turnover number. The EOQ model computes the economic order quantity, which is the economic hiring number in a period. Therefore, the optimal hiring strategy is identified to keep employee inventory stable and to shorten the hiring lead time.

The final step is updating the employee-turnover forecasting model every five years. Because the model is built on historical data, the internal and external factors, as well as the forecasting results, will change after five years. Therefore, it is essential to import the historical employee turnover data into the forecasting model in order to rerun it so that a professional statistician can identify the significant attributes and optimal model.

In conclusion, the employee-turnover model is necessary for the lean management system, which provides support for the system's normal operation and reduces the cost/waste resulting from employee turnover. Implementing this model requires the cooperation of the HR and department managers as well as the HR and statistical staff. As part of a lean

system, this model should be considered as important as Kanban, 5S, and other lean tools. Not limited to manufacturing organizations, the employee-turnover forecasting model could be relevant to service and government organizations.

5.3 Future Work

The employee-turnover model is based on employee records exported from HR's management system. However, some key information, like salary, is not accessible. Nevertheless, many researchers have proved that employee salary is a key factor in employee-turnover (Griffeth et al., 2000). The forecasting model will continue being developed and improved if additional information, like salary, is available.

Also, qualitative interview and survey methods help researchers capture employees' opinions regarding such issues as job satisfaction, leadership, organization commitment, and turnover intention. Those opinions combined with employees' records can build a mixed employee-turnover forecasting model. Regardless of the statistical methods used in building a model, interview and survey methods can be more precisely designed for certain factors among different targeted groups. These factors are managers' greatest concerns. For example, when the turnover rate suddenly increases, interview and survey data can quickly identify the important changing factor, like leadership, organizational structure, or job satisfaction.

Finally, the optimal employee hiring strategy can be explored in industrial and academic areas. Instead of simply applying an EOQ model to determine an optimal hiring number, a more complex optimization model can be developed by considering how to redistribute and retrain current employees to the opening position. This may provide additional value to both academic and industry practitioners.

Bibliography

Abelson, M. A. and Baysinger, B. D. (1984). Optimal and dysfunctional turnover: Toward an organizational level model. *Academy of management Review*, 9(2):331–341. 27

Agresti, A. (1996). *Categorical data analysis*, volume 996. New York: John Wiley & Sons. 84

Alao, D. and Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4(1). 9, 83

Allen, D. G., Bryant, P. C., and Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *The Academy of Management Perspectives*, 24(2):48–64. 80

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide*. Sas Institute. 32, 38, 40, 42, 68

Allway, M. and Corbett, S. (2002). Shifting to lean service: stealing a page from manufacturers' playbooks. *Journal of Organizational Excellence*, 21(2):45–54. 3, 4

Alwan, L. C. and Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1):87–95. 11

Balfour, D. L. and Neff, D. M. (1993). Predicting and managing turnover in human service agencies: A case-study of an organization in crisis. *Public Personnel Management*, 22(3):473–486. 7, 81, 82

Batt, R. (2002). Managing customer services: Human resource practices, quit rates, and sales growth. *Academy of management Journal*, 45(3):587–597. 2

Beng Ang, K., Tee Goh, C., and Chye Koh, H. (1994). An employee turnover prediction model: A study of accountants in singapore. *Asian Review of Accounting*, 2(1):121–138. 83

Berger, J. O. and Chen, M.-H. (1993). Predicting retirement patterns: Prediction for a multinomial distribution with constrained parameter space. *The Statistician*, pages 427–443. 32

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213. 12

Bluedorn, A. C. (1982). A unified model of turnover from organizations. *Human relations*, 35(2):135–153. 7, 81, 82

Bowerman, B. L., O'Connell, R. T., and Koehler, A. B. (2005). *Forecasting, time series, and regression: An applied approach*. South-Western Pub. 17

Bowie, C. and Prothero, D. (1981). Finding causes of seasonal diseases using time series analysis. *International journal of epidemiology*, 10(1):87–92. 20

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (1970). *Time series analysis: forecasting and control*. John Wiley & Sons. 14, 20

Braun, M. and Schweidel, D. A. (2011). Modeling customer lifetimes with multiple causes of churn. *Marketing Science*, 30(5):881–902. 32

Briggs, T. (2014). Survival analysis for predicting employee turnover. http://www.slideshare.net/twbriggs/survivalanalysisforpredictingemployeeturnover. Accessed on 10/29/2015. 32

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75(3):713–737. 31

Brostrm, G. (2015). *eha: Event History Analysis*. R package version 2.4-3. 46, 52

Burke, R. J. (1994). Generation x: Measures, sex and age differences. *Psychological Reports*, 74(2):555–562. 82

Burr, D. (1994). A comparison of certain bootstrap confidence intervals in the cox model. *Journal of the American Statistical Association*, 89(428):1290–1302. 49

Čambál, M., Holková, A., and Lenhardtová, Z. (2011). Basics of the management. 7

Carrión, A., Solano, H., Gamiz, M. L., and Debón, A. (2010). Evaluation of the reliability of a water supply network from right-censored and left-truncated break data. *Water resources management*, 24(12):2917–2935. 32, 37

Chang, H.-Y. and Xi, L. (2009). Employee turnover: a novel prediction solution with effective feature selection. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, number 3. Citeseer. 83

Chatfield, C. (2000). *Time-series forecasting*. CRC Press. 25

Chu, F.-L. (1998). Forecasting tourist arrivals: nonlinear sine wave or arima? *Journal of Travel Research*, 36(3):79–84. 17, 44

Clark, R. L. (2002). Retirement: Early retirement incentives. http://www.encyclopedia.com/topic/Incentive_services.aspx. Accessed: 2015-12-02. 3, 38

Clark, S. C. (2000). Work/family border theory: A new theory of work/family balance. *Human relations*, 53(6):747–770. 91

Claus, E., Risch, N., and Thompson, W. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232. 32

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer. 53

Collett, D. (2015). *Modelling survival data in medical research*. CRC press. 40

Collini, S. A., Guidroz, A. M., and Perez, L. M. (2015). Turnover in health care: the mediating effects of employee engagement. *Journal of nursing management*, 23(2):169–178. 7

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276. 37

D'agostino, R. B., Belanger, A., and D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321. 17

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press. 44, 49

DeLurgio, S. A. (1998). *Forecasting principles and applications*. 14, 17, 20

Detert, J. R., Treviño, L. K., Burris, E. R., and Andiappan, M. (2007). Managerial modes of influence and counterproductivity in organizations: a longitudinal business-unit-level investigation. *Journal of Applied Psychology*, 92(4):993. 2

Feldman, D. C. (1994). The decision to retire early: A review and conceptualization. *Academy of management review*, 19(2):285–311. 77

Ferrara, L. and van Dijk, D. (2014). Forecasting the business cycle. *International Journal of Forecasting*, 30(3):517–519. 7

Glebbeek, A. C. and Bax, E. H. (2004). Is high employee turnover really harmful? an empirical test using company records. *Academy of Management Journal*, 47(2):277 – 286. 2

Griffeth, R. W., Hom, P. W., and Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of management*, 26(3):463–488. 81, 97

Grznar, J., Booth, D. E., and Sebastian, P. (1997). A robust smoothing approach to statistical process control. *Journal of chemical information and computer sciences*, 37(2):241–248. 11

Hamilton, J. D. (2011). Calling recessions in real time. *International Journal of Forecasting*, 27(4):1006–1026. 7

Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., and Pierce, C. A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. *Journal of Management*, 39(3):573–603. 2

Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press. 84

Hanke, J. E., Reitsch, A. G., and Wichern, D. W. (1998). *Business forecasting*. Prentice Hall Upper Saddle River, NJ. 17

Harrell, F. E. (2013). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media. 45

Harvey, A. C. and Todd, P. (1983). Forecasting economic time series with structural and box-jenkins models: A case study. *Journal of Business & Economic Statistics*, 1(4):299–307. 14

Heneman, H., Schwab, D., Fossum, J., and Dyer, L. (1993). *Personnel/human resource management*. The Irwin series in management and the behavioral sciences. Irwin. 6

Holt, B. A. (2011). Development of an optimal replenishment policy for human capital inventory. 32

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression (3rd Edition)*. New York, NY, USA: John Wiley & Sons. 42

Hosmer Jr, D. W. and Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons. 84

Hyndman, R. J. (2014). Cran task view: Time series analysis. http://cran.r-project.org/web/views/TimeSeries.html. Accessed: 2014-05-01. 27

IBM (2013). Hr analytics: Invest in your most valuable asset with hr analytics. http://www-01.ibm.com/software/analytics/solutions/operational-analytics/hr-analytics/. Accessed on 10/29/2015. 28, 32

Ittig, P. T. (1997). A seasonal index for business. *Decision Sciences*, 28(2):335–355. 26

Jackson, T. L. (1996). *Implementing a lean management system*. Productivity press. 4

Jena, R. (2015). An assessment of factors affecting organizational commitment among shift workers in india. *Management: Journal of Contemporary Management Issues*, 20(1):59–77. 91

Jenkins, M. and Paul Thomlinson, R. (1992). Organisational commitment and job satisfaction as predictors of employee turnover intentions. *Management Research News*, 15(10):18–22. 83

Kacmar, K. M., Andrews, M. C., Van Rooy, D. L., Steilberg, R. C., and Cerrone, S. (2006). Sure everyone can be replaced but at what cost? turnover as a predictor of unit-level performance. *Academy of Management Journal*, 49(1):133–144. 1, 2, 3, 28

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons. 32

Kelarijani, S. E. J., Heidarian, A. R., Jamshidi, R., and Khorshidi, M. (2014). Length of service and commitment of nurses in hospitals of social security organization (sso) in tehran. *Caspian journal of internal medicine*, 5(2):94. 91

Khoong, C. (1996). An integrated system framework and analysis methodology for manpower planning. *International Journal of Manpower*, 17(1):26–46. 6, 81

Kilpatrick, J. (2003). Lean principles. *Utah Manufacturing Extension Partnership*, pages 1–5. 4

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media. 37

Kleinbaum, D. G. (1998). Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108. 38

Kochan, T. A. and Useem, M. (1992). *Transforming organizations*. Oxford Univ. Press. 4

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer. 42, 67

Lane, W. R., Looney, S. W., and Wansley, J. W. (1986). An application of the cox proportional hazards model to bank failure. *Journal of Banking & Finance*, 10(4):511–531. 31

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons. 32

LeClere, M. J. (2005). Preface modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, 4(4):5–12. 31

Leonard, B. (2001). Turnover at the top. 1, 81

Lin, Z. and Hui, C. (1999). Should lean replace mass organization systems? a comparative examination from a management coordination perspective. *Journal of International Business Studies*, pages 45–79. 4

Long, J. S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press. 32

Lu, J. (2002). Predicting customer churn in the telecommunications industry: An application of survival analysis modeling using sas. *SAS User Group International (SUGI27) Online Proceedings*, pages 114–27. 31

McShane, S. L. and Von Glinow, M. A. Y. (2003). *Organizational Behavior: Emerging Realities for the Workplace Revolution*. McGraw-Hill/Irwin New York, NY. 82

Meeker, W. Q. and Escobar, L. A. (2014). *Statistical methods for reliability data*. John Wiley & Sons. 32

Mobley, W. H. (1982). Some unanswered questions in turnover and withdrawal research. *Academy of Management Review*, 7(1):111–116. 1

Moeschberger, M. L. and Klein, J. (2003). *Survival analysis: Techniques for censored and truncated data: Statistics for biology and health*. Springer. 32

Moncarz, E., Zhao, J., and Kay, C. (2009). An exploratory study of us lodging properties' organizational practices on employee turnover and retention. *International Journal of Contemporary Hospitality Management*, 21(4):437–458. 96

Moody, R. W. (2000). Going, going, gone by reporting to management on the causes and devastating effects of employee turnover, internal auditors may help to curb the costs. *Internal Auditor*, 57(3):36–41. 81

Morrow, P. C., McElroy, J. C., Laczniak, K. S., and Fenton, J. B. (1999). Using absenteeism and performance to predict employee turnover: Early detection through company records. *Journal of Vocational Behavior*, 55(3):358–374. 7, 82

Mueller, C. W. and Price, J. L. (1989). Some consequences of turnover: A work unit analysis. *Human Relations*, 42(5):389–402. 1, 28

Nagadevara, V., Srinivasan, V., and Valk, R. (2008). Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques. *Research & Practice in Human Resource Management*, 16(2). 82

Ng, S. H., Cram, F., and Jenkins, L. (1991). A proportional hazards regression analysis of employee turnover among nurses in new-zealand. *Human Relations*, 44(12):1313–1330. 9, 82

Organisation for Economic Co-operation and Development (2013). Composite leading indicators(mei). http://stats.oecd.org/index.aspx?queryid=6617. Accessed: 2014-04-30. 10

Pan, W. and Chappell, R. (1998). A nonparametric estimator of survival functions for arbitrarily truncated and censored data. *Lifetime data analysis*, 4(2):187–202. 32

Pankratz, A. (2012). *Forecasting with dynamic regression models*, volume 935. John Wiley & Sons. 14

Paul, A. M. (2012). This is the biggest reason talented young employees quit their jobs. http://mobile.businessinsider.com/why-young-employees-quit-their-jobs-2012-9. Accessed: 2016-02-01. 91

Popoola, S. (2006). Personal factors affecting organizational commitment of records management personnel in nigerian state universities. *Ife Psychologia*, 14(1):183–197. 91

PWC (2015). Dont just collect data. generate insight. http://www.pwc.com/us/en/hr-saratoga.html. Accessed: 2015-11-15. 32

Rainlall, S. (2004). A review of employee motivation theories and their implications for employee retention within organizations. *The journal of American academy of business*, 9:21–26. 30, 31

Rhodes, S. R. (1983). Age-related differences in work attitudes and behavior: A review and conceptual analysis. *Psychological bulletin*, 93(2):328. 91

Sagie, A., Birati, A., and Tziner, A. (2002). Assessing the costs of behavioral and psychological withdrawal: A new model and an empirical illustration. *Applied Psychology*, 51(1):67–89. 1

Selden, S. C. and Moynihan, D. P. (2000). A model of voluntary turnover in state government. *Review of Public Personnel Administration*, 20(2):63–74. 80

Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., and Smith, A. M. (2005). Employee turnover: A neural network solution. *Computers & Operations Research*, 32(10):2635–2651. 7, 9

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, pages 591–611. 17

Shaw, J. D. (2011). Turnover rates and organizational performance review, critique, and research agenda. *Organizational Psychology Review*, 1(3):187–213. 80

Simonoff, J. S. (2013). *Analyzing categorical data*. Springer Science & Business Media. 44

Smith, J. L. (2009). 12 reasons employees leave organizations. http://www.peoriamagazines.com/ibi/2009/dec/12-reasons-employees-leave-organizations. Accessed: 2016-02-01. 91

Staw, B. M. (1980). The consequences of turnover. *Journal of Occupational Behaviour*, pages 253–273. 1, 2, 28

Tableman, M. and Kim, J. S. (2003). *Survival analysis using S: analysis of time-to-event data*. CRC press. 40

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805. 14

Tews, M. J., Stafford, K., and Michel, J. W. (2014). Life happens and people matter: Critical events, constituent attachment, and turnover among part-time hospitality employees. *International Journal of Hospitality Management*, 38:99–105. 7

Thaden, E., Jacobs-Priebe, L., and Evans, S. (2010). Understanding attrition and predicting employment durations of former staff in a public social service organization. *Journal of Social Work*, 10(4):407–435. 7, 82

Tuji, A. (2013). *An Assessment of the Causes of Employee Turnover in Oromia Public Service Organizations*. Thesis. 90

Velicer, W. F. and Fava, J. L. (2003). Time series analysis. *Handbook of psychology*. 9

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, 75(371):609–615. 11

Wang, M. and Shultz, K. S. (2010). Employee retirement: A review and recommendations for future investigation. *Journal of Management*, 36(1):172–206. 31

Warren, H. E. (2008). Forecasting time series within excel. http://web.calstatela.edu/faculty/hwarren/a503/forecast%20time%20series%20within%20Excel.htm. Accessed: 2014-04-30. 27

Weisent, J., Seaver, W., Odoi, A., and Rohrbach, B. (2010). Comparison of three time-series models for predicting campylobacteriosis risk. *Epidemiology and infection*, 138(06):898–906. 17

White, G. L. (1995). Employee turnover: The hidden drain on profits. *HR Focus*, 72(1):15–17. 1

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342. 14

Wright, T. A. and Cropanzano, R. (1998). Emotional exhaustion as a predictor of job performance and voluntary turnover. *Journal of Applied Psychology*, 83(3):486. 7, 82

Yeung, A. K. and Berman, B. (1997). Adding value through human resources: reorienting human resource measurement to drive business performance. 4

Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175. 12

Appendix

Appendix A

Summary of Tables

Table A.1: Time Series Univariate Models for Turnover Data

Model	Variables	Pred. R^2	Holdout R^2	MAPE	Normality	WN
Regression[1]	Additive T+S	0.51	0.57	26.15	No	No
	Additive S	0.45	0.09	22.32	No	No
	Additive T+S+T*S	0.58	0.22	23.41	No	No
	Additive T+S+Intervention[2]	0.72	0.52	22.84	Yes	No
	Multiplicative T.S	0.34	0.55	25.42	Yes	No
	Multiplicative S	0.33	0.50	25.58	Yes	No
	Multiplicative T.S (T*S)	0.30	0.26	29.75	Yes	No
Decomposition[1]	T*C*S	0.65	0.54	17.97	Yes	Yes
WES[1]	Additive T+ Additive S	0.52	0.52	20.65	Yes	Yes
	Additive T+ Multiplicative S	0.54	0.46	20.17	Yes	Yes
	Multiplicative T+ Additive S	0.52	0.40	19.82	Yes	Yes
	Multiplicative T+ Multiplicative S	0.52	0.39	19.77	Yes	Yes
ARIMA	ARIMA(1,0,0)(2,1,0)	0.39	0.33	22.94	No	Yes
	ARIMA(1,0,0)(2,0,0)	0.38	0.38	23.96	Yes	Yes
	ARIMA(0,1,1)(2,1,0)	0.45	0.14	22.96	Yes	Yes
	ARIMA(0,1,1)(2,0,1)	0.42	0.17	22.45	Yes	Yes
	ARIMA(0,0,1)(1,0,1)	0.35	0.30	25.35	Yes	Yes
	ARIMA(0,0,1)(1,0,2)	0.39	0.33	22.96	Yes	No
	ARIMA(1,0,0)(0,1,1)	0.40	0.48	25.20	Yes	Yes
	ARIMA(1,0,1)(2,0,0)	0.43	0.26	21.82	Yes	Yes
	ARIMA(1,0,1)(1,0,2)	0.44	0.20	21.55	Yes	Yes
	ARIMA(1,0,1)(0,1,1)	0.47	0.40	22.89	Yes	Yes
	ARIMA(2,1,0)(0,1,1)	0.41	0.26	25.24	Yes	Yes
	ARIMA(0,0,1)(2,0,0)	0.36	0.37	25.00	Yes	Yes
	ARIMA(1,1,1)(1,0,0)	0.35	0.04	24.20	No	Yes
	ARIMA(1,0,1)(1,1,0)	0.39	0.32	22.45	Yes	Yes
	ARIMA(1,0,1)(1,0,1)	0.41	0.24	22.18	No	Yes
	ARIMA(0,1,1)(2,0,0)	0.40	0.16	22.60	Yes	Yes
	ARIMA(1,1,1)(2,0,0)	0.41	0.20	22.26	Yes	Yes
	ARIMA(1,0,1)(1,0,1)	0.35	0.04	23.87	No	Yes

[1] In Regression, Decomposition and Exponential Smoothing models: T denotes Trend, S denotes seasonality, C denotes cycle, and T*S denotes an interaction term between T and S.

[2] The data is unsmoothed (or outliers are unadjusted) so as to take advantage of time series models that can accommodate interventions.

Table A.2: Dynamic Regression Models with CLI and Its Lags as Cyclical Factor

Model	Pred. R^2	Holdout R^2	MAPE	Normality	WN
Dynamic regression & CLI	0.72	0.56	22.56	Yes	No
Dynamic regression & lag1 CLI	0.72	0.54	22.73	Yes	No
Dynamic regression & lag2 CLI	0.74	0.55	21.87	Yes	No
Dynamic regression & lag3 CLI	0.74	0.55	21.50	Yes	Yes
Dynamic regression & lag4 CLI	0.74	0.56	21.35	Yes	Yes
Dynamic regression & lag5 CLI	0.75	0.57	21.32	Yes	Yes
Dynamic regression & lag6 CLI	0.76	0.58	20.73	Yes	Yes
Dynamic regression & lag7 CLI	0.77	0.59	19.91	Yes	Yes
Dynamic regression & lag8 CLI	0.77	0.59	20.05	Yes	Yes

Table A.3: Decomposition Models with CLI and Its Lags as Cyclical Factor

Model	Pred. R^2	Holdout R^2	MAPE	Normality	WN
Decomposition & CLI	0.65	0.55	17.97	Yes	Yes
Decomposition & lag1 CLI	0.65	0.55	17.97	Yes	Yes
Decomposition & lag2 CLI	0.65	0.55	17.97	Yes	Yes
Decomposition & lag3 CLI	0.65	0.55	17.97	Yes	Yes
Decomposition & lag4 CLI	0.65	0.54	17.97	Yes	Yes
Decomposition & lag5 CLI	0.65	0.54	17.97	Yes	Yes
Decomposition & lag6 CLI	0.65	0.53	17.97	Yes	Yes
Decomposition & lag7 CLI	0.65	0.53	17.97	Yes	Yes
Decomposition & lag8 CLI	0.65	0.53	17.97	Yes	Yes

Table A.4: ARIMA Models with CLI and Its Lags as Cyclical Factor

Model	Pred. R^2	Holdout R^2	MAPE	Normality	WN
ARIMA(1,0,1)(0,1,1) & CLI	0.37	0.41	22.77	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag1 CLI	0.37	0.41	22.73	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag2 CLI	0.37	0.41	22.78	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag3 CLI	0.37	0.41	22.85	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag4 CLI	0.36	0.40	22.91	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag5 CLI	0.36	0.40	22.98	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag6 CLI	0.36	0.40	23.04	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag7 CLI	0.36	0.40	23.09	Yes	Yes
ARIMA(1,0,1)(0,1,1) & lag8 CLI	0.36	0.40	23.12	Yes	Yes

Vita

Xiaojuan Zhu was born in Tianjin, China on March 30, 1980, daughter of Honggang Zhu & Jun Liu. She completed her high school in 1999 and enrolled at Management Department of Hefei University of Technology, major in Information System and Information Management. Xiaojuan graduated with her bachelor degree in fall 2003 and joined Tianjin Mitsumi Electric Company as a production planning engineer to gain practical experience in the field. After gained three-year working experience, she was admitted into the Education Economics and Management master program from University of Science and Technology Beijing in September 2007. After Xiaojuan had accomplished her master study, she came to U.S. to unite her family in July 2009. Xiaojuan started her Ph.D. in Industrial Engineering at University of Tennessee, Knoxville in August 2011, focusing on the area of employee turnover forecasting. She severed as a graduate research assistant in the university during her Ph.D. study and completed her Ph.D. in summer 2016.