# SURVIVING A CIVIL WAR: EXPANDING THE SCOPE OF SURVIVAL ANALYSIS IN POLITICAL SCIENCE

by

Andrew B. Whetten

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

| | |
|---|---|
| John R. Stevens, Ph.D.<br>Major Professor | Richard Cutler, Ph.D.<br>Committee Member |
| Damon Cann, Ph.D.<br>Committee Member | Laurens H. Smith, Ph.D.<br>Interim Vice President for Research and Dean of the School of Graduate Studies |

UTAH STATE UNIVERSITY
Logan, Utah

2018

# Abstract

Surviving a Civil War: Expanding the Scope of Survival Analysis in Political Science

by

Andrew B. Whetten, Master of Science

Utah State University, 2018

Major Professor: John R. Stevens, Ph.D.
Department: Mathematics and Statistics

Survival Analysis in the context of Political Science is frequently used to study the duration of agreements, political party influence, wars, senator term lengths, etc. This paper surveys a collection of methods implemented on a modified version of the Power-Sharing Event Dataset (which documents civil war peace agreement durations in the Post-Cold War era) in order to identify the research questions that are optimally addressed by each method. A primary comparison will be made between a Cox Proportional Hazards Model using some advanced capabilities in the glmnet package, a Survival Random Forest Model, and a Survival SVM. En route to this comparison, issues including Cox Model variable selection using the LASSO, identification of clusters using Hierarchal Clustering, and discretizing the response for Classification Analysis will be discussed. The results of the analysis will be used to justify the need and accessibility of the Survival Random Forest algorithm as an additional tool for survival analysis.

(117 pages)

# Public Abstract

Surviving a Civil War: Expanding the Scope of Survival Analysis in Political Science

by

Andrew B. Whetten, Master of Science

Utah State University, 2018

Major Professor: John R. Stevens, Ph.D.
Department: Mathematics and Statistics

Survival Analysis in the context of Political Science is frequently used to study the duration of agreements, political party influence, wars, senator term lengths, etc. This paper surveys a collection of methods implemented on a modified version of the Power-Sharing Event Dataset (which documents civil war peace agreement durations in the Post-Cold War era) in order to identify the research questions that are optimally addressed by each method. A primary comparison will be made between a Cox Proportional Hazards Model using some advanced capabilities in the glmnet package, a Survival Random Forest Model, and a Survival SVM. En route to this comparison, issues including Cox Model variable selection using the LASSO, identification of clusters using Hierarchal Clustering, and discretizing the response for Classification Analysis will be discussed. The results of the analysis will be used to justify the need and accessibility of the Survival Random Forest algorithm as an additional tool for survival analysis.

(117 pages)

To my wife and my son who will be joining the pack October 2018.

## Acknowledgments

I am grateful for the support and guidance of Dr. Stevens who has been an understanding advisor through my graduate degree. I am indebted to my parents who instilled in me a love of learning and hard work. Most importantly, I am thankful for my wife for her support, passion for education, and willingness to remain poor graduate students for a few more years.

Andrew B. Whetten

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Across all observation and experimentation found in the social sciences, statistical methods are frequently inspired by or adapted to currently arising research questions. Even though many methods cannot be mirrored identically from previous analyses, a vast quantity of potentially useful methods are left unused. In the discipline of political science, event history models are frequently constructed from time-to-event data. The abundantly utilized method of analysis is the Cox Proportional Hazards Model [1]. Various other methods and expansions to the Cox Model have been utilized in social science event history modeling, including parametric models, multiple-spell duration models, discrete-time models, time dependent covariate models, and accelerated failure time models [2]. In some scenarios, these methods may have significant drawbacks that inhibit researchers from effectively answering some research questions. The fundamental purpose of this paper is to articulate the applicability of more advanced event history and classification methods in the social sciences. Emphasis will be placed on more modern survival models. In the present case, a dataset monitoring the duration of civil war peace agreements during the post-Cold War era will be utilized to exemplify the advantages of using an arsenal of statistical methods in the analytic research which allows one to gain broader insight and greater understanding of the benefits and drawbacks of various methods.

## 1.1 Motivating Example: The Power Sharing Event Dataset

The Power-Sharing Event Dataset (PSED) documents the power-sharing arrangements and civil-war recurrence between government-rebel dyads from 1988 to 2006 [3]. The observational unit that has been used in previous analysis is an individual peace agreement during the first five years of its implementation. Binary covariates can be used to evaluate

the qualitative features of a respective peace agreement [4]. These include evaluations of the promise and implementation of economic, military, territorial, and political power-sharing, where (1, 0) denotes a (presence, absence) of a power sharing element. These are referred to as the generalized power-sharing variables. These have been constructed from detailed power-sharing variables that are within each category of power-sharing. The presence of several specific peace agreement features are identified by a unique binary quality – these features each have two covariates representing the promise and then implementation of a respective feature as well. There are 39 candidate covariates in this initial dataset. Not all of these can be used at once since strong collinearity will exist between the detailed and generalized power-sharing variables. For the vast majority of the paper, the generalized power-sharing variables will be used exclusively.

In his master's work in Utah State's Political Science Department [4], Chong Chen contributed additional peace agreement evaluation criteria to the PSED by adding the binary covariates Mediated Design and Mediated Implementation. The power-sharing covariates have also been summarized into more general power-sharing covariates by Chen that identify if at least one element of a respective power-sharing type is present in a peace agreement. Chen's thesis focused on the interaction between the presence of mediated design and implementation using a Cox Proportional Hazards Model. In this type of analysis, all peace agreements that had durations exceeding the 5 year scope of the study are considered right censored [4]. Unlike many examples found in clinical trials, there were no "dropout patients" since peace agreements could be followed until the end of study without ethical concerns. Peace agreements that have lasted longer than 5 years will be referred to as successful peace agreements. Note that since the PSED is an observational study (rather than a designed experiment), conclusions are limited to association, instead of causation. Additionally, the data collected on peace agreement implementation may not be complete. For this reason, some variables may not be predictively valuable. For the purposes of this paper, this will not cause an issue for the demonstration of statistical methods. Variables with no information (such as all zeros for a specific peace agreement characteristic) were

screened and disqualified from the analysis.

The dataset comprises 79 peace agreements, as unique observational units, that are valid for a classical event history model that monitors for the time-to-failure of a peace agreement. The original dataset boasts 353 observations that detail critical events that occur during the peace process, resulting in multiple event-types occurring to the same 79 peace agreements. These observations in the context of most methods over-inflate the dataset's true sample size. The various types of events that can occur before a failure-event or right censoring of a peace agreement may be each useful in subsequent individual analyses on respective event types, or a competing risks model is required. Another critical element of this analysis is the clustered nature of these 79 peace agreements. As an example, 10 of these peace agreements involve the government of Chad.

The response variable, duration, identifies the length of time from signing to violation of a peace agreement. Any peace agreements that lasted 5 years (1826-1827 days), are right censored observations. In the collection of this dataset, 9 observations had peace agreement violations that did not have events times known to an exact day. Some of these observations had event times rounded to the beginning of a week or end of a month. One extreme case had an event censored until the beginning of the peace agreement. This would suggest that interval censored methods [5] may be worth considering even though it is an uncommon occurrence in Political Science. Variables used in the main analysis of the PSED are summarized in Table 1.1.

## 1.2 Methodology

The goal of this paper is to apply various statistical techniques to Chen's modified PSED [4] in order to identify the benefits and deficiencies of each method, and through this process, determine which questions may be answered most effectively. The approach will consist of presenting an array of candidate survival and classification methods. In all of the analyses that will be applied, the abbreviated 79 observation dataset will be used, because a univariate response variable is the most common set up for time-to-event data. Competing Risks models are available for some of the methods that will be presented and

are worth examining in future analysis of the PSED. The focus of the analysis will be to examine the model fit, predictive accuracies, and the interpretation and importance of the predictor variables.

For some of the methods, the detailed power-sharing variables from the original PSED are analyzed instead of using Chen's generalized power-sharing variables. These will be included in Appendix A.1 and A.2.

Table 1.1: Power-Sharing Event Dataset variables descriptions: Detailed power-sharing variables are not included in this table.

| Variable Name | Description |
|---|---|
| med design | Presence/absence of mediated design of an agreement by a third party |
| med imp | Presence/absence of mediated implementation of an agreement by a third party |
| inter | The interaction between mediated design and implementation |
| intDY | The conflict intensity preceding an agreement |
| durationDY | The duration of conflict preceding an agreement |
| multireb | Presence/absence multiple rebel party signatories |
| unpko | Presence/absence of UN peacekeepers |
| ps political | Presence/absence of political power sharing |
| ps military | Presence/absence of military power sharing |
| ps economic | Presence/absence of economic power sharing |
| ps territory | Presence/absence of territorial power sharing |
| ppsPROM IMP | Promise and implementation of political power sharing present/absent |
| epsPROM IMP | Promise and implementation of economic power sharing present/absent |
| mpsPROM IMP | Promise and implementation of military power sharing present/absent |
| tpsPROM IMP | Promise and implementation of territorial power sharing present/absent |

# Chapter 2

# Summary of Candidate Classification and Survival Methods Implemented

Several methods could be applied to the PSED, with different focuses and research questions. The methods considered here include unsupervised learning, survival based methods, and supervised learning.

## 2.1 Unsupervised Learning

### 2.1.1 Hierarchal Cluster Analysis

An inspection of clustering structure is not always intuitive and unsupervised learning techniques are an effective approach for gaining further insight. Hierarchal Clustering involves stepwise partitioning of observations and clusters respective to distance in the p-dimensional space determined by the covariates, where p refers to the number of covariates in the model. The survival time is not considered in this method since clustering is only examined across the covariates; this also prevents the need to account for a censored response variable. This is most commonly performed in using a "bottom-up" approach known as Agglomerative Clustering where the initial amount of clusters is equal to the amount of observations. The process of merging clusters is repeated until all observations are grouped into a single final cluster and the merging of observations and clusters can be visualized using a dendrogram. Divisive Clustering is performed in the opposite direction where an all-encompassing single cluster is constructed and the clusters are repeatedly divided from parent clusters [6].

In the context of the PSED, there are several ways that clustering may be considered to occur. For the purpose of this paper, only clustering by location is examined. While

inspecting the resulting dendrogram(s), resulting clusters will be inspected for patterns denoted by the peace agreements location by country. Peace agreements that have highly similar attributes will be closer together in the p-dimensional space constructed by the peace agreement characteristics. Hence, they will be grouped together early in the dendrogram.

The distance between two observations $x_i = (x_{i1}, x_{i1}, \ldots, x_{ip})$ and $x_j = (x_{j1}, x_{j1}, \ldots, x_{jp})$ for p numeric variables is calculated using the Minkowski Metric

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g)^{1/g} \tag{2.1}$$

The value of g determines the the type of paraxial distance where the commonly used Euclidean distance is acquired using $g = 2$. Numeric variables should be standardized before using this Metric or a weighted distance metric is required.

Calculating p-dimensional distance between two observations for binary and categorical variables requires the use of a 2x2 contingency table that counts matching presences, matching absences, and disagreeing attributes between two observations. For a symmetric binary attribute, where both outcomes are considered equally valuable, a simple matching coefficient can be used to determine the difference using the equation

$$d(x_i, x_j) = \frac{b + c}{a + b + c + d}, \tag{2.2}$$

where $a$ is the number of attributes that equal 1 for both objects, $b$ and $c$ are numbers of disagreeing binary attributes and $d$ is the number of attributes that equal zero for both objects [7].

Frequently, observations are characterized by mixed attribute types, as in the PSED where both binary and continuous variables exist in the model, and the distance between observations is calculated using the generalized dissimilarity equation for p attributes

$$d(x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_{ij}^{(n)} d_{ij}^{n}}{\sum_{n=1}^{p} \delta_{ij}^{(n)}}, \tag{2.3}$$

where indictor $\delta_{ij}^{(n)}$ equals zero if a missing value is present for the attribute in either object i or j. "The contribution of attribute n to the distance between the two objects $d_{ij}^{(n)}$ is computed according to its type" [7].

The distance between clusters can be defined in various ways. Four frequently used methods are described below:

1. Single Linkage identifies the distance between two clusters as the shortest distance between single observations in each cluster. This method is also referred to as the nearest-neighbor methods.

2. Complete Linkage identifies the distance between two clusters as the farthest distance between single observations in respective clusters. This method is also referred to as the furthest-neighbor method.

3. Average Linkage identifies the distance between two clusters as the average distance from observations of one cluster to observations in the other cluster.

4. Ward's (Minimum Variance) Method combines clusters that minimize the total increase of within-cluster variance.

## 2.2 Survival Methods

Survival methods, frequently referred to as event history methods, are characterized by a response variable that denotes time-to-event. The event in the context of Biostatistics typically refer to death, acquisition of illness, recovery of patient etc. Political scientists have adapted many of these methods to examine the duration of a current circumstance based on the history preceding the event [2].

A critical feature of survival methods is accounting for observations in the risk set. The risk set involves all observations that are still at risk of experiencing an event. Depending on the nature of the observational unit, an event occurrence may permanently inhibit observations from remaining in the risk set such as death of the observational unit. Other

event types, such as election of political leaders, may allow observations to return to the risk set in the future.

Observations can be kept in the risk set after a termination of a study, late entry, or dropout of an observation unit by accounting for censoring. Right censoring identifies a final time that an observation was known to exist in its present state, but retains the observation in the risk set. Right censored observations only contribute information of survival until the censored event time. Left censoring identifies an unknown initial time of the present state of the observational unit, and only enter the risk set at the beginning of the study. Similarly, left censored observations only contribute information from the start of the study onward [2].

### 2.2.1 Cox Proportional Hazards Model and LASSO Extension

The Cox Regression Model utilizes the hazard rate defined by the function $h(t) = f(t)/S(t)$ where f(t) represents the density function of "the [instantaneous] unconditional failure rate of event occurrences" and S(t) represents "the proportion of units surviving beyond [a given time] t" [2]. It was established in 1972 that the hazard function can remain arbitrary by establishing the proportional hazards assumption given by the equation

$$h(t) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p), \tag{2.4}$$

where parameter estimates are acquired maximizing the partial likelihood function and log partial likelihood function

$$L(\beta) = \frac{\prod_{r \in D} exp(\beta^T x^{j_r})}{\sum j \in R_r exp(\beta^T x^j)} \tag{2.5}$$

$$l(\beta) = logL(\beta) \tag{2.6}$$

The LASSO (Least Absolute Shrinkage Selection Operator)was initially developed in the linear regression context when Tibshirani "proposed minimization of the RSS, subject

to a constraint of the form $\sum|\beta_j| \leq s$" [8]. Since the Cox Proportional Hazards (Cox PH) estimates are most commonly estimated using the partial likelihood, the shrinkage constraint is applied to the log partial likelihood which shrinks initial solutions to zero where $\sum|\hat{\beta}_j| < s$. In his paper, Tibshirani recalls that the advantage of this constraint is the frequent occurrence of solution coefficients becoming exactly zero, unlike Ridge Regression. This sacrifices some model stability found in the Ridging Algorithm for increased interpretability. The LASSO criterion to estimate Cox Model parameters is the following [9]:

$$\hat{\beta} = argmin(l(\beta)) \quad \text{subject to} \quad \sum|\beta_j| \leq s \tag{2.7}$$

### 2.2.2 Interval Censored Log-Rank and Cox Proportional Hazards Models

In some cases it may be critical to account for interval censoring in methods that have this capability. Unlike left and right censoring, where censoring is performed on observation that did not experience an event during the time period of the study, interval censoring addresses missing response information on observations that did experience the event. This occurs when an observation does not have a precise failure time, but rather an upper and lower bound where the event is known to occur.

For the nonparametric log rank method, interval censoring is accounted for in the likelihood function

$$L = \prod_{i}^{n}[S(L_i) - S(R_i)], \tag{2.8}$$

where the difference of the survival curves $S(t) = Pr(T_i > t)$ for the left and right endpoints of the interval $(L_i, R_i]$ for all $i$ observations for which a observation event-time $T_i$ is observed. For more information about the resulting nonparametric maximum likelihood estimator (NPMLE) refer to Peto 1973 [10] and So *et al.* 2010 [11].

For interval censoring in the Cox Proportional Hazards Model, the data is structured as $(A_i, X_i), i = 1...n$ where $A_i = (L_i, R_i]$ is the interval where an event time was known to occur and $X_i$ is the vector of covariates for observation i [12]. Interval censoring is accounted

for similarly to the log rank method by using the likelihood function

$$L = \prod_{i}^{n} [G(L_i|x_i) - G(R_i|x_i)], \tag{2.9}$$

where the difference of the hazard curves $G(t) = Pr(T > t|X = x)$ for the left and right endpoints of the interval $(L_i, R_i]$ for all $i$ observations for which an observation event-time is observed between $[T_{j-1}, T_j]$ where $T_1...T_m$ are the follow-up times in the study. Finkelstein [12] establishes that the proportional hazards assumption defines the probability of an observation surviving longer than $T_j$ with its regression or (covariate vector) $x_i$ as

$$G(T_j|x_i) = [G(T_j)]^{exp(\beta x_i)}. \tag{2.10}$$

Clustering can be accounted for in a Shared Frailty Model [5]. The resulting Cox proportional hazards equation is expressed as

$$h(t) = Z_i h_0(t) exp(\beta x_1 + \beta x_2 + \ldots + \beta x_p), \tag{2.11}$$

where $Z_i$ is the frailty penalty corresponding to the $i^{th}$ cluster.

### 2.2.3 The Survival Random Forest Algorithm

In 2008 Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone and Michael S. Lauer published a paper that extended the application of the Random Forest algorithm to the context of survival analysis [13]. In their paper, it is argued that the use of Survival Random Forest (SRF) is of great value since "survival data are commonly analyzed using methods that rely on restrictive assumptions such as proportional hazards" [13]. Using SRF prevents the potential need to transform nonlinear variables, use specialized base functions, perform stepwise regression or variable selection procedures, and identify high order interactions. These problems are addressed automatically in Random Forest procedures.

The SRF procedure is defined by Ishwaran to follow "the prescription laid out by Breiman" where the response must be taken into account when growing a survival tree. In

this process, the splitting criterion uses the censoring information and the time-to-event to split by a predictor variable at a given node in a way that maximizes the survival difference between the two resuting groups. This is different than survival forest models that had been developed in previous years which include single Survival Trees, Relative Risk Forests, and weight RF regression analysis for right censored data [13]. The algorithm for SRF can be broadly summarized as the following list of 5 steps [14]:

1. Draw *ntree* (number of trees) bootstrap samples from the original data.
2. Grow a tree for each bootstrapped data set. At each node of the tree randomly select *mtry* predictors for splitting on. Split on a predictor using a survival splitting criterion. A node is split on that predictor which maximizes survival differences across daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than *nodesize* unique deaths.
4. Calculate an ensemble cumulative hazard estimate by combining information from the *ntree* trees. One estimate for each individual in the data is calculated.
5. Compute an out-of-bag (OOB) error rate for the ensemble derived using the first b trees, where b = 1, ... , *ntree*.

There are 4 conventionally used splitting criterion put forth by Ishwaran: log-rank splitting, conservation of events splitting, log-rank score splitting, and approximate log-rank splitting [14].

To assess the predictive ability of a SRF, concordance error rate is used. To calculate this error rate all possible pairs of observations are examined and pairs that do not meet "permissible criterion" are removed. Permissible criterion for calculating a concordance error rate requires that a pair of observations must not have a right censored observation with a recorded duration shorter than a non-censored observation. This criterion requires that all ties be eliminated except for ties where one observation is right censored when another observation has a recorded failure time. Then the OOB estimates of each observation

within each remaining pair are compared to examine whether or not the actual survival times agree with the compared OOB estimates.

It is important to note that the SRF procedure has been shown to outperform in terms of concordance error rates the Cox and censored Random Forest regression methods. This will not be demonstrated in this paper for Random Forest Regression. Comparing the SRF concordance and the Cox Proportional Hazard concordance would be a meaningful assessment of how well the SRF procedure performs in comparison to the more conventional approach [13].

Similar to the other Random Forest procedures, variable importance can be evaluated. This is performed in a slightly different manner for survival trees. Variable importance is calculated by comparing optimally split in-bag trees at a specific covariate to randomly split in-bag trees at the same specific covariate. This is performed in the following manner. Each OOB observation for each tree is run down their respective survival tree. Whenever a split is encountered at a given variable, x, assign a daughter node randomly. The cumulative hazard function is then calculated and averaged for all such trees where a variable x has been randomized. The variable importance is then calculated by subtracting the prediction error from the original/optimal ensemble from the prediction error rate for the new ensemble that randomizes at the x-split [13].

### 2.2.4 Survival Support Vector Machines

A recent alternative to the Cox Model is to implement support vector machines (SVMs) in the context of censored time-to-event data. This approach at a foundational level is worth consideration due to "their good theoretic foundations and high classification accuracy" [15]. SVM classification (summarized more generally in Section 2.3.7) supposes a response variable with two or more classes, such as a recorded failure or no-failure of a peace agreement, and classifies observations by constructing a separating hyperplane between the classes with an optimized margin. The margin defines the shortest distance from an observation to the separating hyperplane. Data points that define this shortest distance to the hyperplane are called support vectors. Observations may be incorrectly classified if two classes are

not separable, and these observations are penalized for being misclassified by using slack variables [15].

This approach has been extended to the regression setting and ultimately to the context of censored survival analysis. This was accomplished by three different approaches: regression, ranking, and a hybrid combination of both ranking and regression. All three of these methods aim to maximize the concordance index for permissible pairs of observations. The ranking approach, the earliest and the most intuitive and computationally efficient approach, converts the problem into a ranking calculation by characterizing an observation's performance relative to the other observations. Evaluating concordance indices is inherently a prediction evaluation based on rank since pairs of observations are only recorded to have longer, shorter, or equal predicted survival times rather than what magnitude of difference are a pair of predictions [15, 16].

## 2.3 Supervised Learning

The purpose of supervised learning methods in the context of this paper is to attempt to construct a meaningful and predictively accurate classification model using a simplified response variable. This is done by converting the response into two groups, the failed peace agreements and the peace agreements that lasted longer than 5 years. This can be done since no observations are right censored before the end of study. The predictors used for these methods are restricted to the generalized power-sharing variables and all other variables that do not relate to power-sharing. Generally simplifying the response results in a loss of information and is not ideal if a continuous response is recorded. The objective of this implementation is to attempt to construct a model that provides comparable results to the survival analysis methods, and to discuss what research questions can be answered with such a model.

### 2.3.1 Linear and Quadratic Discriminant Analysis

Linear and Quadratic Discriminant Analysis both involve using a specific geometric classification surface to split the observations into two or more predefined groups by min-

imizing the amount of misclassified observations. These methods require the multivariate Gaussian assumption for all continuous predictor variables. This would require a log transformation of the durationDY variable (representing the duration of conflict), which is newly labeled as LNdurationDY. New or cross validated observations are assigned to a classification group by calculating the Mahalanobis Distance between an observation and a given group mean. The Mahalanobis Distance accounts for the spread of the groups in the p-dimensional space using the covariance of a given group [17]. These methods are differentiated by the type of classification surface that is used and the resulting manner by which the Mahalanobis Distance is calculated.

Linear Discriminant Analysis uses p-dimensional linear decision boundaries. This choice of surface is typically optimized by rough equivalence of group covariance matrices. The covariance matrices are pooled and the resulting equation for the Mahalanobis distance becomes

$$D^2(x, \overline{x}_j) = (x - \overline{x}_j) S_p^{-1} (x - \overline{x}_j), \tag{2.12}$$

where $S_p^{-1}$ is the pooled covariance matrix for all groups.

Quadratic Discriminant Analysis uses n-dimensional quadratic decision boundaries. Using a quadratic decision boundary can reduce error rates at decision boundaries of groups with distinctly different dimensionality. Covariance matrices in this method are not pooled and the Mahalanobis Distance is expressed by

$$D^2(x, \overline{x}_j) = (x - \overline{x}_j) S_j^{-1} (x - \overline{x}_j), \tag{2.13}$$

where $S_j^{-1}$ is the covariance matrix corresponding to group $j$ .

### 2.3.2 k-Nearest Neighbor Method (kNN)

The kNN method classifies new or cross validated observations by voting them into a group determined by the majority of neighboring observation class-type. Neighboring observations are defined as the "closest" observations to the observation of interest. The

distance of the nearest neighbors is most commonly determined using the Mahalanobis Distance [17], and the class-type of the neighbors is defined by the response variable. The simple tuning parameter that is adjusted multiple times is k, the quantity of neighbors permitted for voting. Generally, a smaller quantity of neighbors produces better cross-validated error rates and larger numbers of neighbors tend to include observations that are more distant from the observation of interest, as identified by the Mahalanobis Distance. Observations that are relatively far away are generally expected to belong to different class-types and they can potentially confound the voted results from the nearer neighbors that have stronger similarities to the observation of interest.

### 2.3.3 Generalized Linear Model: Logistic Regression

Gerenalized linear models (GLMs) form a broad class of models involving a response $y_i$ that belong to the linear-exponential family. Included in this class of models are several frequently used distributions that are all interrelated such as the normal, exponential, gamma, binomial, Poisson, Inverse Gaussian etc. Three main components are found in all GLM's:

1. The random component is the probability distribution of the response variable that belongs to some member of the linear-exponential family, with mean $\mu$.

2. The systematic component is the combination of linear predictors that can be represented by

$$\eta_i = \beta^T X = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.14}$$

   where $X = (x_1 \ldots x_p)$ are the predictors variables with the corresponding effect parameters $\beta_1 \ldots \beta_p$ .

3. The link function g is the determined relationship between the linear combination of the predictor variables and the response. More concisely, it is the specified link between the mean of the random component and systematic component given by

$$g(\mu) = \eta. \tag{2.15}$$

In statistical software packages for GLM procedures, one can perform a binary logistic regression by specifying the random component to follow the binomial distribution $B(n, \pi)$ where the mean $\pi$ is the probability of observing a 1 or "success." A common link function is the log-odds or Logit link defined by

$$log(\frac{\pi}{1-\pi}) = \eta. \tag{2.16}$$

There is some added flexibility in using GLM packages, such as the glmmer package in R or the GLIMMIX procedure in SAS, to perform logistic regression. Both of these packages allow for generalized linear mixed models which properly account for random and fixed effects in the same model. As an example, these packages are more accommodating of accounting for blocking factors, which can be a useful tool for examining clustered data.

### 2.3.4 Classification/Decision Trees

Decision Trees are the simplest of the tree-based methods. Since right-censored observations prevent the option of using Regression Trees, only Classification trees will be described here. Decision Trees are constructed by stratifying the predictor variables using an optimized split based on maximizing purity of the resulting daughter nodes. A Classification tree is used to predict a qualitative result. Predictions for test observations are determined by the most frequently occurring class from training observations with similar predictor variable characteristics.

Class proportions at each node are important for identifying the optimized node purity at a given split. Class proportions are calculated at each node and they are typically displayed in finalized visualizations of the tree. There are multiple ways to calculate purity. The Gini index is a frequently used measure of node purity that is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.17}$$

where $\hat{p}_{mk}$ is the proportion of observations from the $m^{th}$ predictor region and the $k_{th}$ class. Note that as the purity, or proportion in one class, increases, the Gini index decreases towards zero [17].

### 2.3.5 Random Forest Algorithm

Random Forest classification and regression uses bootstrap samples of the data to grow many trees in a process called bagging. Decision trees are prone to relatively high model variance. Growing several trees with bootstrap samples of the data and averaging the predictions from all trees grown, will decrease the overall variance of the statistical learning method. Out-of-bag (OOB) observations are used to evaluate the prediction accuracy of the model. This is accomplished by using the OOB observations from each respective tree and acquiring predictions for each of these observations using the bootstrap sampled training model [17].

The attribute of the Random Forest algorithm that distinguishes it from a standard Bagging procedure that can be implemented with any statistical learning method is the restriction of the number of predictors allowed at each node split within a tree. At each split, a random sample of $m$ predictors is selected as candidate variables for the optimal split. This is understood to decorrelate the trees within a Random Forest model since strong predictor variables may potentially take a consistent precedence over other variables in the Bagging procedure. Highly correlated trees do not reduce overall model variability and in the case of having dominant predictors, random sampling of candidate predictors will increase the diversity of the bootstrap sampled trees. The predictions from averaging of more diverse trees will be less variable and more reliable [17].

### 2.3.6 Boosted Trees

Boosting or Gradient Boosting in the context of decision trees is a sequential procedure

of growing trees from previously grown trees. A tree is fitted to the residuals from the previous tree. Excluding the first tree, this is done to all trees as opposed to fitting the model to the response. This new decision tree is then added to the classification or regression function as a new tree and the residuals are recalculated. This is repeated for many trees. Trees are generally smaller tuned to an optimal size. Tree size along with several other parameters are tuned in order to optimize predictive accuracy [17].

### 2.3.7 Support Vector Machines

Support Vector Machines (SVM) is a Classification Algorithm that constructs a hyperplane composed of an optimized set of vectors, known as support vectors, that define a complex boundary between groups of observations. Two parameters are used to tune the SVM typically referred to as C and $\gamma$ that determine the cost of misclassification and the magnitude of influence of training observations. SVMs can also be extended to regression. For the PSED, only right censored regression can be performed using Survival SVMs (summarized in Section 2.2.4).

The parameter C specifies the strictness of the boundary. Larger C values have a stricter margin which allows the SVM to construct a more complicated boundary with larger amounts of support vectors, which decreases the misclassification rate at the expense of over-fitting to the training observations. Lower C-values soften the boundary and permit higher misclassification of training observations in order to decrease the bias of the model.

The $\gamma$ parameter adjusts the distance of influence that observations are permitted to have. Larger $\gamma$ values overfit the model and smaller $\gamma$ values constrain the model and inhibit the model from mapping the shape of the data.

SVM's also have the ability to map the data into higher dimensions using a kernel function where a simpler hyperplane may possibly be used to divide the data more effectively.

# Chapter 3

# Results

## 3.1 Using Hierarchal Analysis to Identify Clusters

Using the agnes function in the cluster package in R [18], the agglomerative clustering coefficient can be used to determine which linkage methods are more optimal. The agglomerative coefficient characterizes the strength of clustering in the data. In this example all cluster coefficients are high and similar across all four methods as shown in Table 3.1. To examine the clustered structure that has been detected, dendrograms can be constructed for each linkage method.

Table 3.1: Examination of Clustering Coefficient for Various Linkages.

| Linkage Method | Average | Single | Complete | Ward |
|---|---|---|---|---|
| Clustering Coefficient | 0.9979 | 0.9966 | 0.9984 | 0.9994 |

In the original PSED, the most interpretable features to help assist in accounting for and characterizing clustering are country, region, and dyadic party. Location and dyadic party tend to limit the ability to effectively account for clustering since many observations would belong in unique clusters which causes issues with the convergence of parameter estimates. The issue with the region variable is that it appears to be continental which may not provide interpretable clusters. As examples from the original regions defined from the PSED, the United Kingdom and Serbia are allocated to the same region, and the peace agreements that occurred in Afghanistan and the Philippines are allocated to the same region. The regions that were ultimately used were recoded through consultation and an examination of agglomerative dendrograms.

A Ward's Method Dendrogram is constructed in Fig. 3.1 in attempt to detect meaningful clustering structure. A larger distance line on the y-axis represents less similar clusters. There are five or six plausible clusters that can be identified. As mentioned before, the

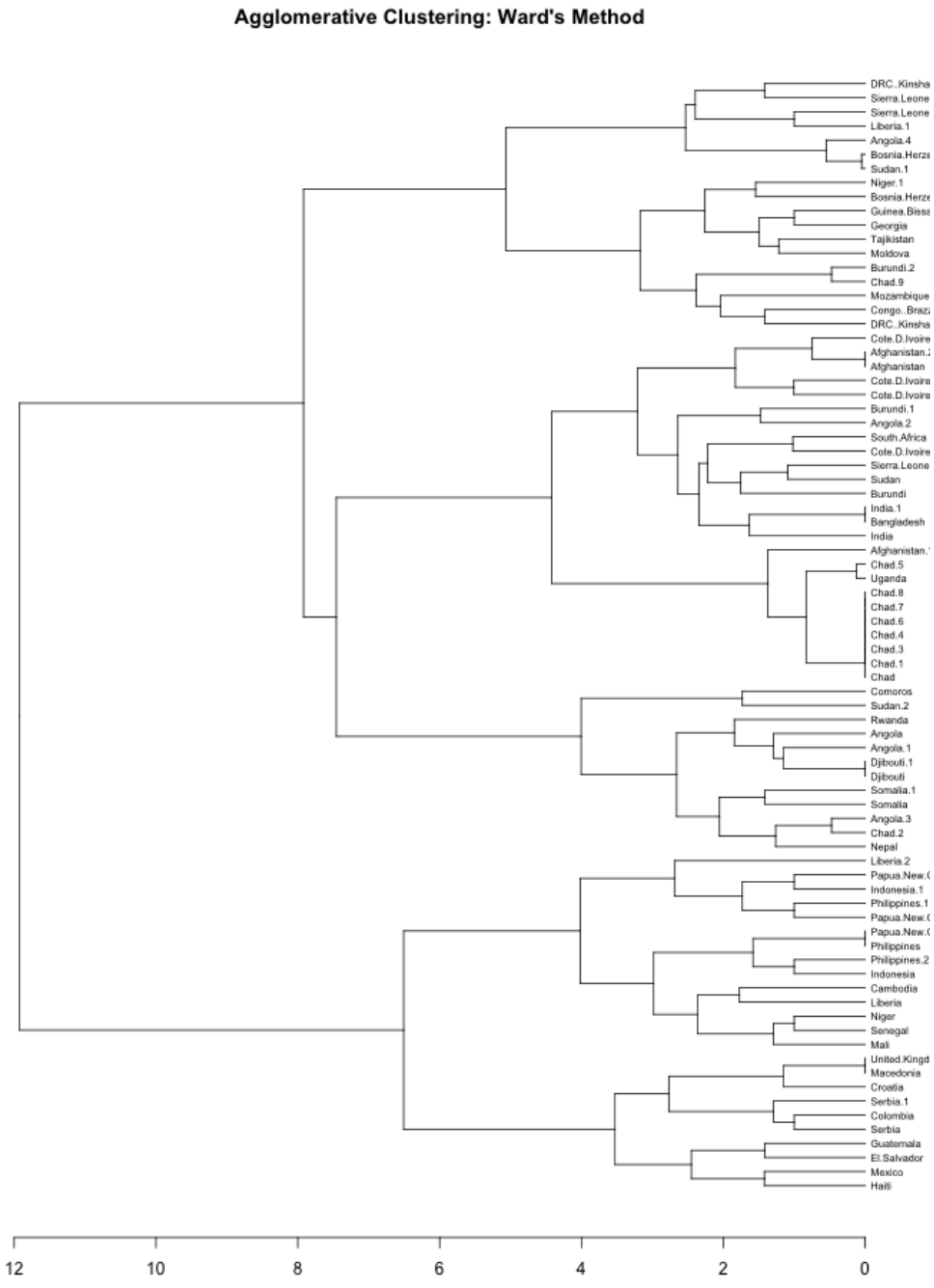


Fig. 3.1: Agglomerative Clustering using Ward's Minimum Variance Method.

interpretation of these clusters may not always be meaningful, so it is worth inspecting each of the clusters more meticulously. In the bottom cluster of Fig. 3.1, all of the South and Central American countries were grouped together with three of the European countries. In the remaining candidate clusters, there is some heterogeneity that confounds that interpretation of the clusters marginally. Some examples include a cluster that groups the Ivory Coast with Afghanistan, Sudan, and India. Despite some less meaningful placements of peace agreements, this clustering method created clusters than had a tendency to group regionally. When assigning peace agreements to clusters for future analysis, regional grouping was still utilized with some modifications resulting from the cluster analysis. This was not heavily emphasized in this paper. In future analyses, a closer inspection of clustering by region and peace agreement characteristics could be implemented.

After constructing a standard agglomerative dendrogram, a heat map can be constructed to inspect for any clustering patterns of peace agreements across the selected predictors. Heat maps can assist in identifying why some clusters may have less interpretable clusters. This was performed using complete-linkage clustering. Note that the the complete linkage dendrogram in Fig. 3.2 has roughly 5-6 distinct clusters with some notable differences in tree shape. A few important features are worth noting. The conflict duration has a significant grouping effect on peace agreements that proceeded lengthy conflicts (refer to dark purple region in Fig. 3.2 corresponding to an notably unique cluster). Recall that conflict duration is the only continuous variable in the PSED. This is likely preventing these peace agreements from finding clusters based on other peace agreement features.

Mediated design, mediated implementation, and the corresponding interaction term also played an important role in clustering of peace agreements. The top cluster and the second from the bottom of Fig. 3.2 are both characterized by having peace agreements with pure or almost purely mediated peace agreements. The differences between these clusters appear in the other power-sharing criterion.

In the third from the top cluster in Fig. 3.2, most of the Chad peace agreements had been appropriately collected. Note that this cluster has peace agreements that have

territorial power-sharing and essentially no other characteristics. The Chad agreement labeled "Chad.9" did not make it in this cluster, but instead found itself in the top cluster with the mediation criterion. This can be noticed for several other countries that had multiple peace agreements over the duration of the observational study. Generally speaking there is some evidence that countries and regions have similar tendencies in the composition of peace agreements.

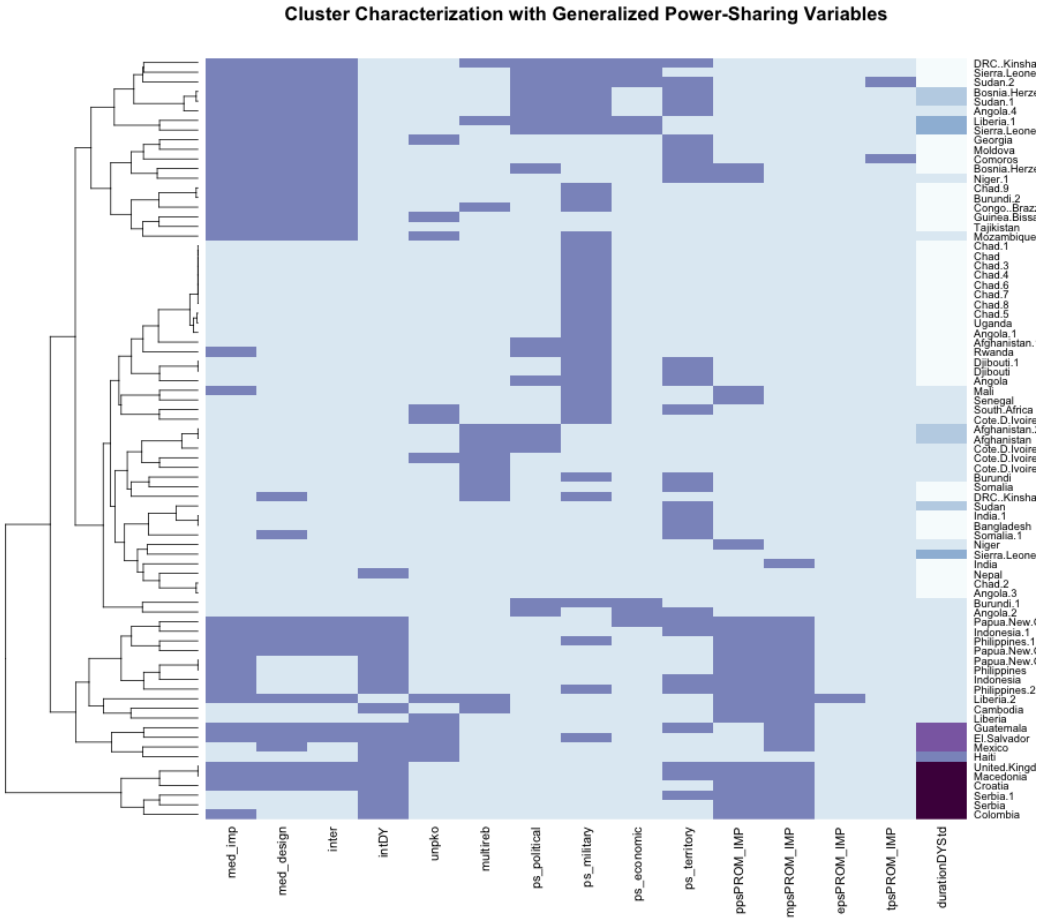


Fig. 3.2: Heat Map Dendrogram using generalized power-sharing variables. Clusters are strongly dependent on interaction of mediated design and implementation.

## 3.2 Survival Analysis Approach

The results of this section will be reported using the Cox Model as a baseline for comparison, then LASSO Selection will be performed on the Cox Model, then a Survival Random Forest Model and Survival SVM will be constructed. Each method will be performed for two initial sets of variables: the generalized power-sharing variables and the detailed power-sharing variables.

### 3.2.1 The Cox Baseline Approach

An initial screening for collinearity was performed on the PSED with only the generalized power-sharing variables included in the model. Table 3.2 shows the variance inflation factors (VIF) for the variables included in the baseline model. There may be collinearity between some of the variables. This does not always affect predictive capabilities of the model, but the interpretability of the coefficients is jeopardized. If collinearity is detected, then eliminating some of the variables is strongly encouraged and may even improve the statistical significance of some of the remaining variables.

Table 3.2: Variance Inflation for Generalized Power-Sharing Variables

| PA Criterion | VIF |
|---|---|
| med imp | 4.761 |
| med design | 7.244 |
| inter | 10.867 |
| durationDY | 1.781 |
| intDY | 4.595 |
| unpko | 1.362 |
| multireb | 1.242 |
| ps political | 1.821 |
| ps military | 1.399 |
| ps economic | 1.554 |
| ps territory | 1.182 |
| ppsPROM IMP | 2.738 |
| mpsPROM IMP | 5.156 |
| epsPROM IMP | 1.505 |
| tpsPROM IMP | 1.160 |

The result for two Cox Models are shown in Table 3.3 for the generalized set of power sharing variables and the detailed set of power sharing variables. Note that of the 15

predictor variables, at least seven of them have highly insignificant p-values which identifies that there is insufficient evidence to identify a significant parameter estimate effect on the response. There are multiple reasons that variable selection is an important consideration.

Some variables in the model have minimal or zero information. This was identified by a visual screening through some of the highly insignificant variables, and some variables were identified to have zeros for all observations. This implies that either no information for these variables was able to be recorded or that none of the peace agreements incorporated this criterion. Some the generalized power-sharing predictor variables are highly zero-inflated, meaning that few or none of the 79 peace agreements implemented some criteria. These predictors provide little or no information about the effect of these features being implemented and act as noise in the data. Removing these variables by some process would be advised as well, and if they are not removed before the analysis, it would be ideal for them to be removed through a variable selection procedure. This is especially critical in higher dimensional data sets where, unlike the PSED, evaluating each predictor variable individually is not feasible.

Table 3.3: Coefficients for Baseline Cox Model

| PA Criterion | Est. | Hazard Ratio | SE(Est.) | z | p |
|---|---|---|---|---|---|
| med design | 1.52 | 4.58 | 0.864 | 1.76 | 0.0785 |
| med imp | 1.03 | 2.81 | 0.727 | 1.42 | 0.1553 |
| inter | -2.95 | 0.05.21 | 1.15 | -2.58 | 0.0099 |
| intDY | -1.16 | 0.313 | 0.728 | -1.60 | 0.1102 |
| LNdurationDY | -0.891 | 0.410 | 0.299 | -2.98 | 0.0029 |
| unpko | 0.507 | 1.66 | 0.764 | 0.66 | 0.5071 |
| multireb | 1.30 | 3.68 | 0.533 | 2.44 | 0.0146 |
| ps political | 0.674 | 1.96 | 0.615 | 1.10 | 0.2731 |
| ps military | -0.921 | 0.398 | 0.474 | -1.94 | 0.0519 |
| ps economic | -0.690 | 0.502 | 1.19 | -0.58 | 0.5615 |
| ps territory | -1.82 | 0.162 | 0.617 | -2.95 | 0.0032 |
| ppsPROM IMP | 0.662 | 1.94 | 0.613 | 1.08 | 0.2808 |
| mpsPROM IMP | 0.900 | 2.46 | 0.662 | 1.36 | 0.1743 |
| epsPROM IMP | -1.96 | 0.142 | 1.54 | -1.27 | 0.2055 |
| tpsPROM IMP | -17.0 | 4.10e-08 | 5220.0 | 0.00 | 0.9974 |

### 3.2.2 LASSO Cox Regression

A LASSO Cox Regression was constructed using the glmnet package [19]. Two plots were created to examine coefficient stability and predictive ability, and then a table of coefficients is provided that summarizes the remaining variables in the Cox Model.

Fig. 3.3 shows the coefficient estimates as variables enter the model under the LASSO constraint. It can be seen that most of the 15 predictor variables have coefficients that stabilize quickly while three to five of them show no evidence of converging. This is a confirmation that some collinearity exists in the model and the parameter estimates from the baseline are not appropriately interpretable.

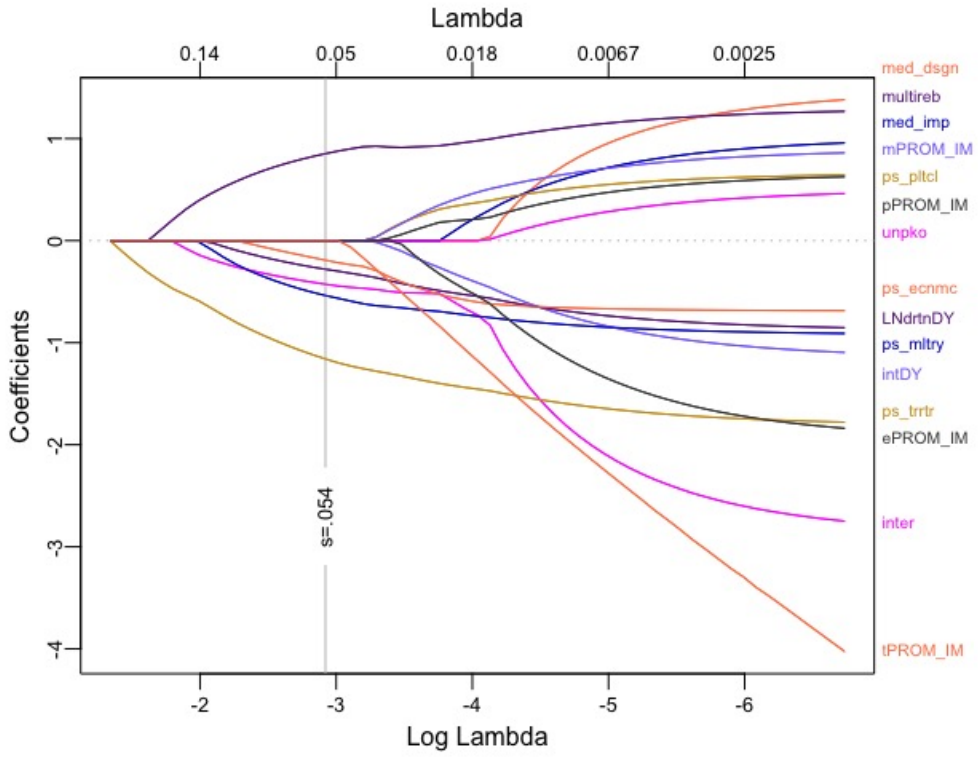


Fig. 3.3: Screening for stability of parameter coefficient estimates for generalized power-sharing variables. $\lambda_{min} = s = 0.054$ provided in plot. Six variables are retained at this point.

Cross validation was implemented using the cv.glmnet function for several $\lambda$ values.

These values correspond to a quantity of predictor variables that have entered the model. The partial likelihood deviance is calculated to assess the predictive fit of the various models. The deviance is minimized for $\lambda = 0.06489$, corresponding to six predictor variables remaining in the model. A typical convention is to also report and consider the simplest model within 1 standard error. The 1-SE model has a $\lambda = 0.14991$ corresponding to 4 predictor variables in the model. Referring to Fig. 3.4, models with 7 to 4 variables are all reasonable candidate model since the Deviance is only marginally jeopardized.

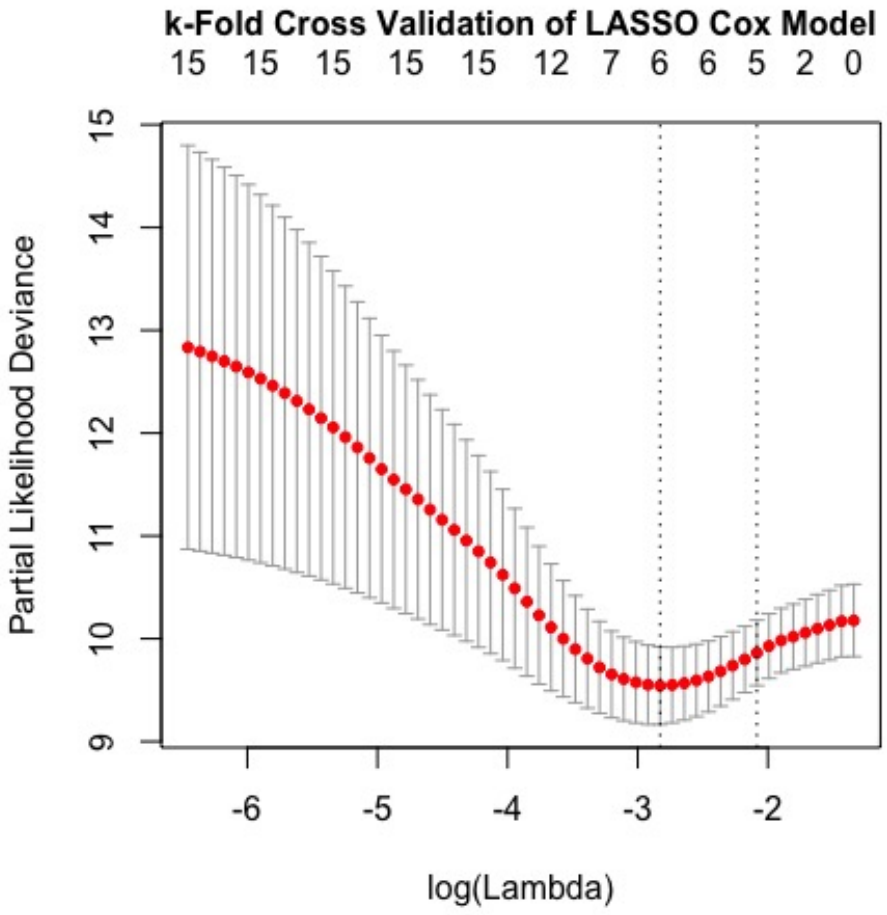


Fig. 3.4: k-Fold cross validation of LASSO increasing Lambda constraint. The number of predictors corresponding to the log(Lamba) value is shown on the upper axis of the plot.

Table 3.4 reports the parameter estimates for the LASSO selected model for the

$\lambda_{min}$ value. A problem that is important to note is that the LASSO model selected the interaction term between mediated design and mediated implementation without selecting the terms individually. This is something that would actually be expected assuming that the LASSO procedure eliminated the collinearity in the model. Referring back to the VIF's in Table 3.2, it can be identified that these two mediation variables and their interaction term have high VIF values. It is reasonable that an interaction term would be collinear with its comprising factors. In Table 3.4, mediated design and mediated implementation were added back into the model and the adjusted LASSO estimates and corresponding p-values are provided. Since clustering was identified, a robust estimate of variance was implemented using the cluster() in the coxph() function of the survival package [20] statement to account for clustering, and the robust p-values and standard errors are provided for comparison.

Table 3.4: Variable Selection using LASSO Cox Model

| | | | Unclustered | | Clustered | |
|---|---|---|---|---|---|---|
| PA Criterion | $\lambda min$ Est. | Adj. Est. | P | SE | Robust P | Robust SE |
| med design | . | 1.2138 | (0.1212) | 0.7832 | (.0902) | 0.7163 |
| med imp | . | 1.1596 | (0.0582) | 0.6122 | (0.0801) | 0.6626 |
| inter | -0.635 | -2.8717 | (0.0098) | 1.1118 | (0.0200) | 1.2349 |
| LNdurationDY | -0.517 | -0.6703 | (0.0019) | 0.2163 | (0.0057) | 0.2423 |
| multireb | 1.150 | 1.3698 | (0.0017) | 0.4368 | (0.0054) | 0.4925 |
| ps military | -0.831 | -0.8883 | (0.0260) | 0.3991 | (0.0101) | 0.3452 |
| ps economic | -0.663 | -0.5046 | (0.6419) | 1.0851 | (0.5235) | 0.7910 |
| ps territory | -1.652 | -1.7614 | (0.0025) | 0.5827 | (<0.001) | 0.2171 |

There are three frequently used residual diagnostic plots for Cox Models. The Schoenfield Residuals are used to detect violations of the proportional hazards assumption, the Martingale residuals are used to detect nonlinearity in the model, and the Deviance Residuals are used to detect influential points. The Schoenfield Residuals are shown across all LASSO selected predictor variables in Fig. 3.5. Most of the predictor variables do not violate the proportional hazards assumption, but the military and territorial power-sharing variables show a clear violation since the confidence bands for most of the predictors included the value zero. The coxzph() function of the survival package [20] was used to identify the significance of the violation. This is done by correlating the residuals with time to test

for independence and reporting the chi-squared statistics for the global and individual predictors. Territorial power-sharing had the only marginally significant p-value, $p = 0.0515$. The global evaluation of the proportional hazards assumption was insignificant, $p = 0.8136$. There is insufficient evidence to reject the proportional hazards assumption. For martingale and deviance residual plots refer to Appendix A.3

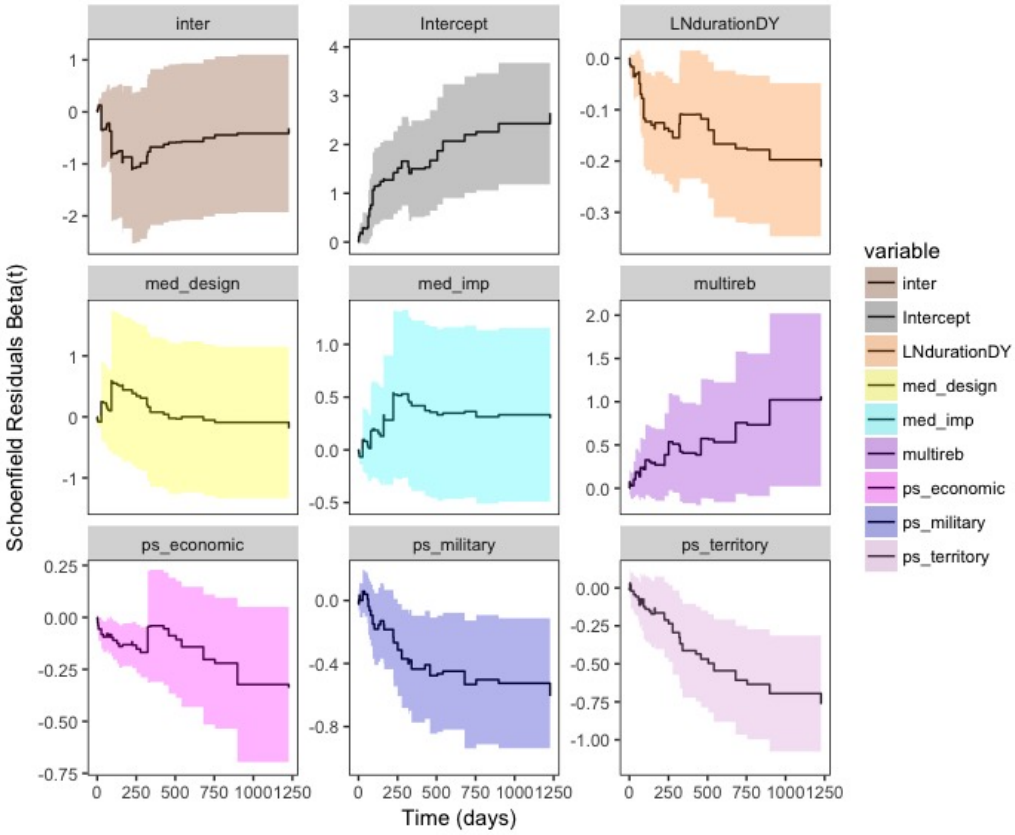


Fig. 3.5: Visual verification of the proportional hazard assumption.

Survival curves were then plotted in Fig. 3.6 by mediation strata, which corresponds to the combinations of mediated design (med design) and mediated implementation (med imp). Confidence bands were plotted for each curve to assist in detecting significant differences between strata. The ability to identify significant differences between strata is limited by the wide confidence bands on the "only mediated implementation" and "only mediated design" categories. Since there is overlap for all strata there is weak evidence to suggest detectable differences between the mediation strata.

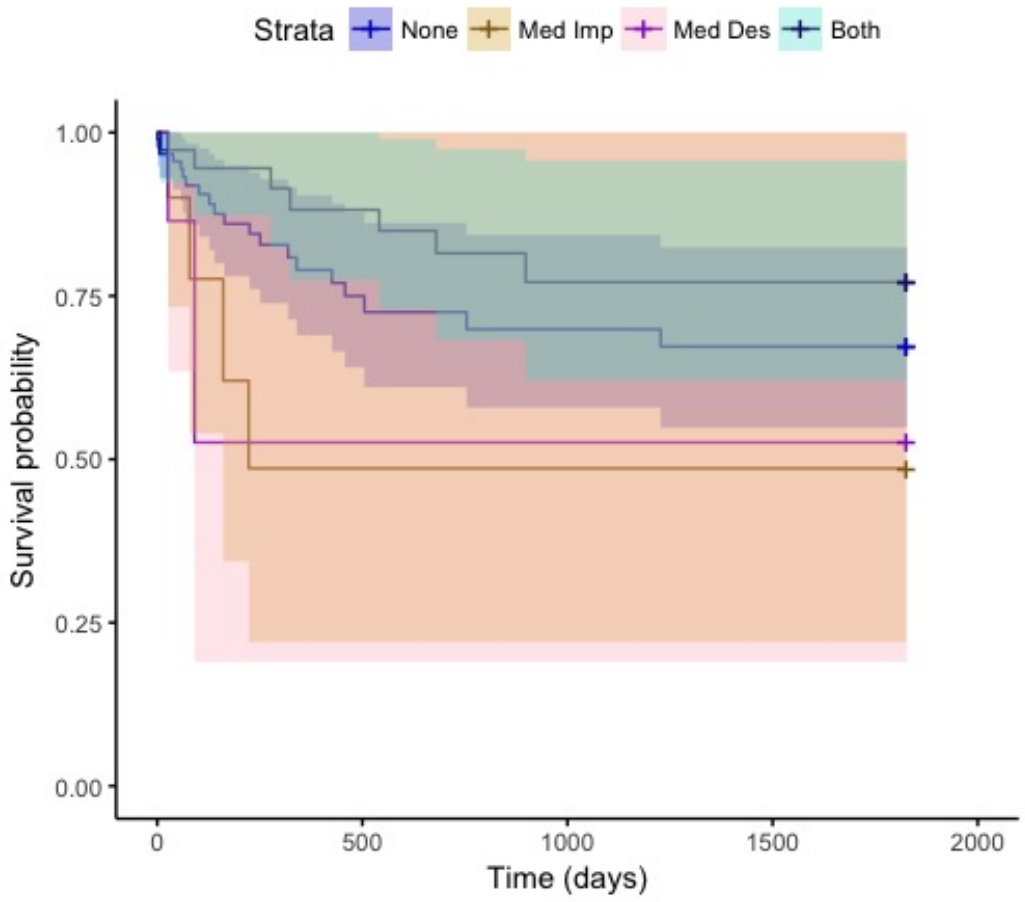


Fig. 3.6: Cox Survival curves by mediation strata using the robust estimate of variance to account for clustering.

### 3.2.3 Interval Censored Log-Rank and Cox Frailty Models

Since nine peace agreements had appropriate information to identify an interval where an event observation was known to occur, interval censored log ranks and Cox Regression methods can be implemented. This is not a conventional occurrence in Political Science, but results are provided for comparison. Even though standard log rank survival curves were not examined for comparison, this method is demonstrated for the interval censored version of the PSED data. Refer to Fig. 3.7 for the log-rank survival curves by mediation strata. Note that these results are consistent with the LASSO Cox PH in Fig 3.6 model that only accounted for right censoring.

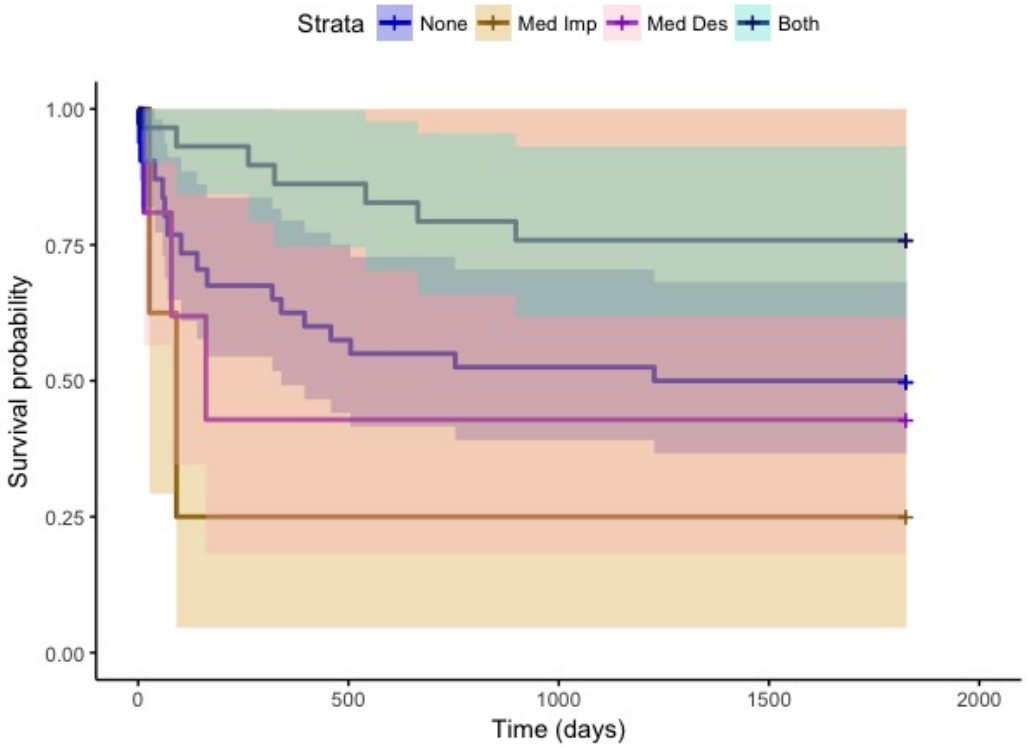


Fig. 3.7: Log-rank Survival Curves accounting for interval censoring.

An interval censored Cox Proportional Hazards model was constructed for comparison with the LASSO selected model. Coefficients are reported in Table 3.5. Comparing these results to the LASSO Cox Model in Table 3.4, it can be observed that some of the estimates are almost unchanged, such as for the mediation variables, while others were more affected.

However, sign values of parameter estimates are consistent and the significance levels of the parameter estimates are relatively consistent.

Table 3.5: Summary of Cox interval censored frailty model for LASSO selected predictors variables.

| PA Criterion | coef | exp(coef) | SE coef | z | p |
|---|---|---|---|---|---|
| cluster(regionAdj) | -0.218 | 0.802 | 0.198 | -1.101 | 0.271 |
| med design | 1.434 | 4.195 | 0.800 | 1.793 | 0.073 |
| med imp | 1.409 | 4.092 | 0.674 | 2.090 | 0.037 |
| inter | -3.651 | 0.026 | 1.199 | -3.046 | 0.002 |
| LNdurationDY | -0.463 | 0.630 | 0.163 | -2.831 | 0.005 |
| multireb | 0.925 | 2.522 | 0.460 | 2.010 | 0.044 |
| ps military | -1.341 | 0.262 | 0.493 | -2.718 | 0.007 |
| ps economic | -0.157 | 0.855 | 1.106 | -0.142 | 0.887 |
| ps territory | -1.542 | 0.214 | 0.581 | -2.652 | 0.008 |

### 3.2.4 Survival Random Forest

A Survival Random Forest (SRF) model can be constructed using the rfsrc() function in the "randomForestSRC" package in R [14]. The default number of trees, 1000, was used in this analysis. The Conservation of Events and Log-Rank splitting criterion were both used and there was minimal difference in the results. The Conservation of Events splitting criterion was used since the Log-Rank has a tendency to favor continuous predictor variables [14].

There are extensive plotting capabilities for the SRF models in the "ggRandomForest" package [21]. For an initial inspection of the model, all of the predicted survival curves can be plotted for each peace agreement. This is shown in Fig. 3.8. Survival probability curves are colored by observed versus unobserved failure. Recall that an unobserved failure in the PSED corresponds to a peace agreement that lasted an unknown duration longer than 5 years. It can be identified that there are generally more blue curves with higher survival probabilities which corresponds to successful peace agreements having a higher prediction probability of success. Visualizing the SRF assists in quickly verifying that the model is generally predicting failed peace agreements to have less chance of survival than successful peace agreements. The extent that the model accurately predicts will be determined using

concordance error.

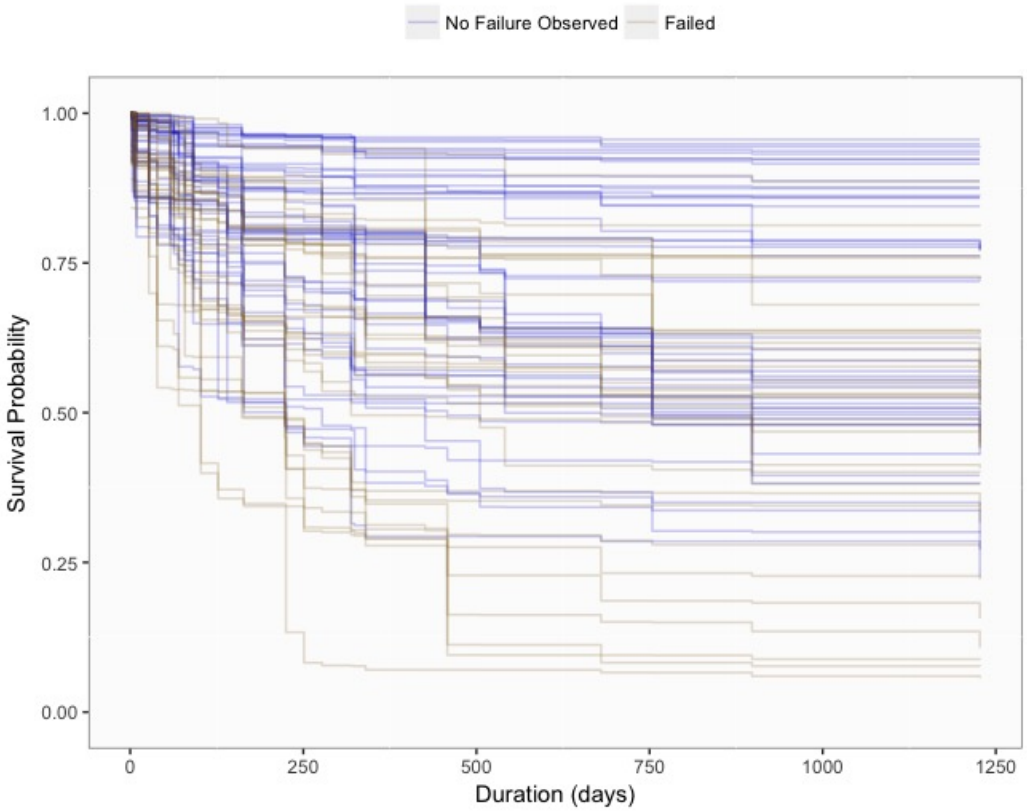


Fig. 3.8: Predicted survival curves for all 79 peace agreements. It can be generally observed that the SRF model is predicting failed peace agreements to have lower survival probability as time increases. Plotting survival probability plots such for all observations with averaging across strata generally creates plots that have lots of noise. This is primarily useful as a initial inspection plot for a survival model.

Fig. 3.9 provides a simple bar plot of variable importance. Note that all of the variables in the adjusted LASSO selected Cox model round out the top eight most important variables in the SRF model. An important feature of the variable importance is the splitting criterion. The Conservation of Events splitting criterion was used since the log-rank splitting criterion tends to favor continuous predictor variables. Since this criterion was used, conflict duration "LNdurationDY" is ranked lower than other splitting criterion choices. Several variables have either negative or approximately zero importance in predicting the lasting success of peace agreements. Negative variable importance identifies that there is a decrease in the prediction performance rather than an increase. This identifies that the variable does not

influence prediction.

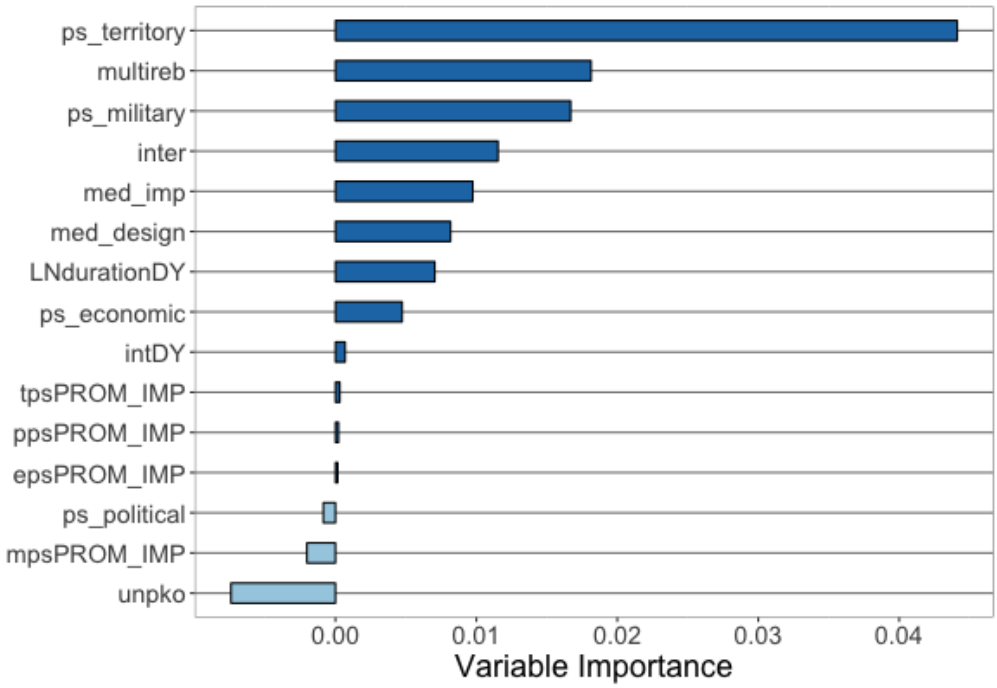


Fig. 3.9: Variable Importance for SRF with Conservation of Events splitting criterion.

Plotting survival curves by strata allows a crisp comparison to the other survival methods. In Fig. 3.10 the averaged survival curves across mediation strata are plotted with the same color coding as the Cox and log-rank models. One of the most striking features of this visualization is the tightness of the confidence bands. In this model, a significant difference can be detected between the peace agreements with completed mediation versus those with partial or no mediation.

In order to quantify the relationship of individuals on the duration of a respective peace agreement, partial dependence plots are constructed for each predictor variable Fig. 3.11. These plots represent the relationships between predictor variable values and the results from hundreds to thousands of bootstrap sampled survival trees. For categorical predictor variables, partial dependence plots are conventionally visualized using box plots. These plots do not allow for more exact relationship quantification in the manner that parametric and semi-parametric models calculate relationship between the predictors and the response,

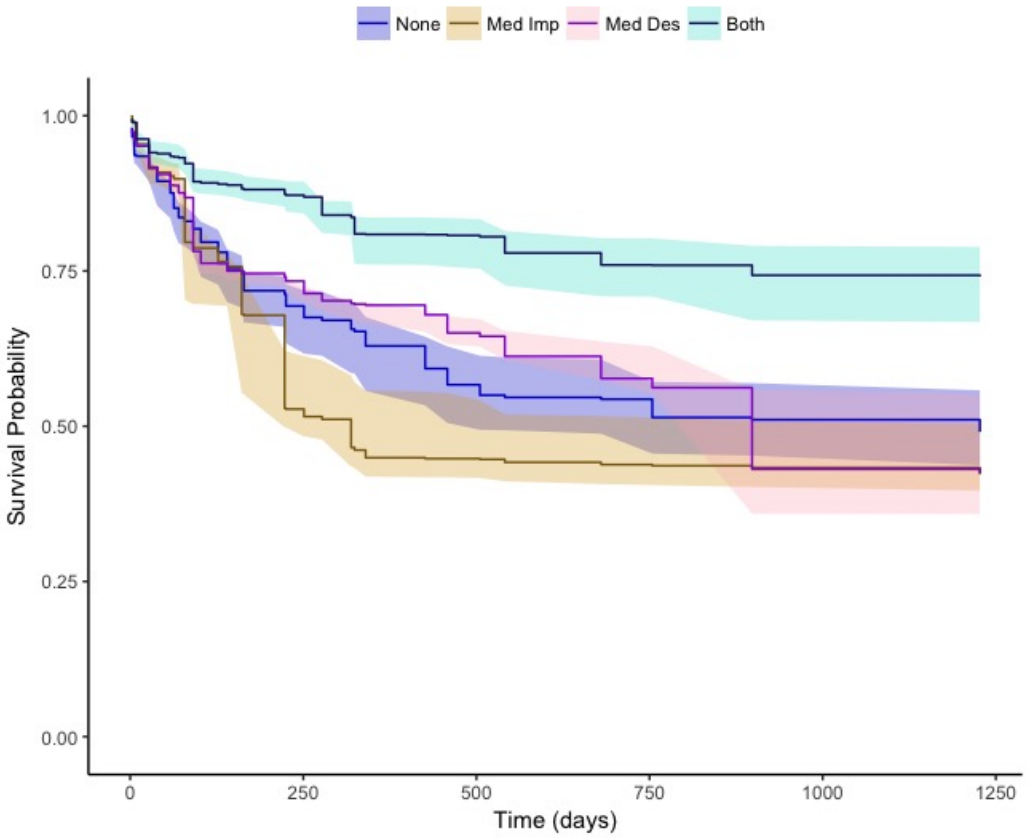


Fig. 3.10: SRF Survival Curves by Mediation Strata. All generalized power-sharing and mediation variables were incorporated in the model. The default of 1000 survival trees were grown to construct a generic SRF model with the conservation of events splitting criterion.

but very similar information can be gathered from partial dependence plots. Partial dependence plots for survival data label the predicted response as "mortality," referring to the predicted number of failed peace agreements by levels of the categorical predictor variables. The constructed box plots for a predictor variable characterize the failure rate across the bootstrap samples for all peace agreements in a given category. To examine all 4 mediation strata in a partial dependence plot, the "inter" variable was recoded from a binary variable into a categorical variable with 4 categories which identify the type of mediation present in a given strata. The "inter" variable's new numbering in Fig. 3.11 corresponds to the following labeling: 0 = no mediation, 1 = mediated implementation present, 2 = mediated design present, 3 = both mediated design and implementation present.

The variables worth noting are territorial power-sharing, the mediation interaction term, and multiple rebel signatories. For the territorial power-sharing criterion, it can explicitly be identified that having a ps-territory value of 1, representing some form of territorial power sharing existing in an agreement tends to reduce the mortality/failure of peace agreements. Even though there is overlap between the two box plots' mortality rates, the 1 category has 3 quartiles of observed bootstrap sampled mortality within the lowest quartile of the 0 category.

Similarly this can be identified in the mediation interaction plot, where the inter=3 corresponds to having both mediated design and mediated implementation. This plot is in agreement with the comparison of survival probabilities by mediation strata in Fig. 3.10. In the multi rebel signatories partial dependence plot, the 1 category has notably higher peace agreement mortality. This is compelling evidence that an increased number of rebel parties involved in a single peace agreement decreases the probability of a peace agreement having a lasting duration.

For many of the less important predictor variables, the lack of a strong relationship was confirmed by observing similar box plots for respective categories.

The preceding conflict duration variable, LNdurationDY, has an interesting relationship visualized in its partial dependence plot. Shorter and longer conflict durations showed

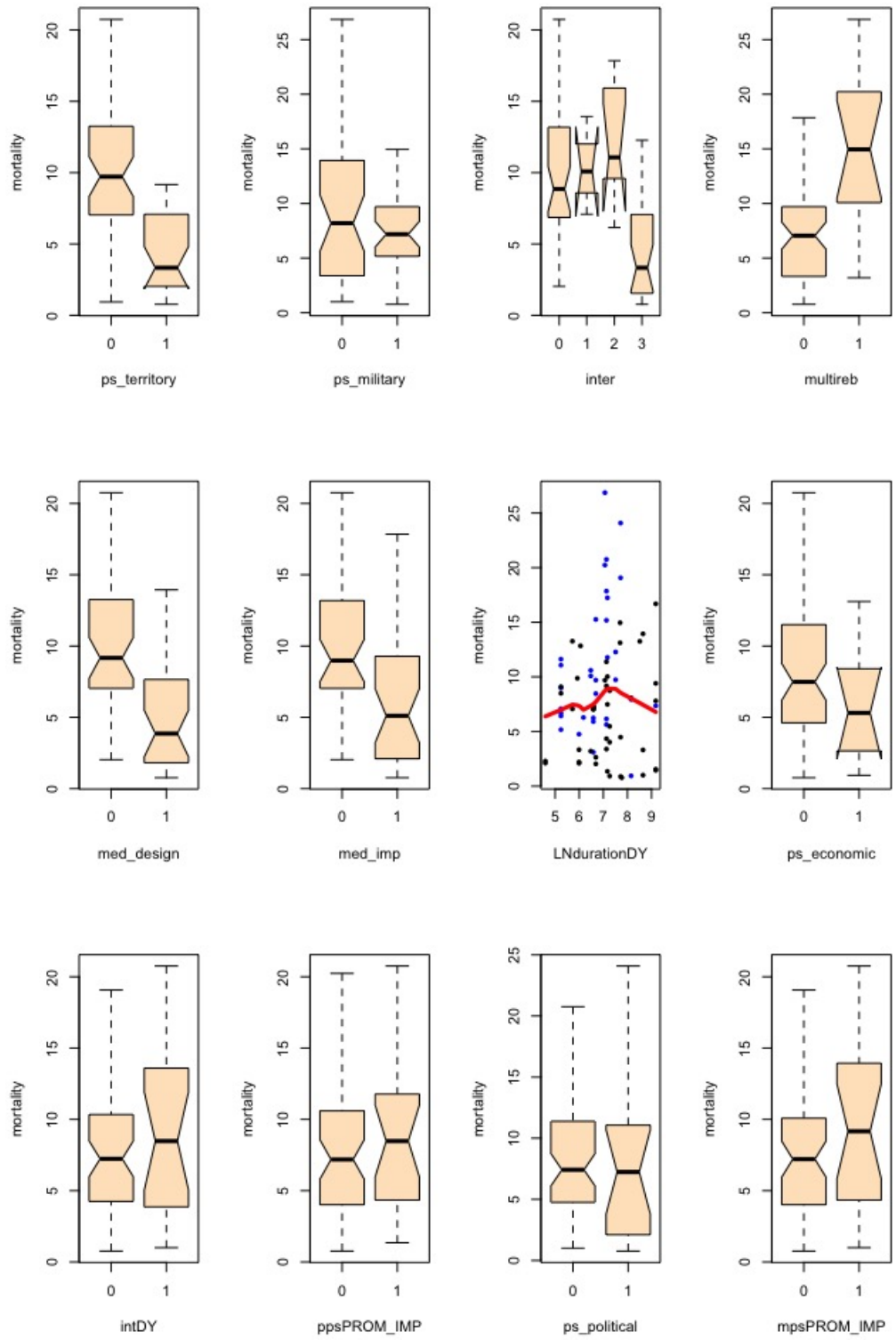


Fig. 3.11: Partial dependence plots for twelve of the fifteen predictor variables utilized. General trends in peace agreement mortality for the various agreement criterion categories. Variables are not ordered by variable importance.

mild tendencies of having lower peace agreement mortality while moderate length conflict durations had stronger tendencies to have higher peace agreement mortality. It could potentially be inferred that longer conflict durations have opposing parties that are more willing to commit to lasting peace, and shorter conflict durations may have conflicts that were easier to resolve, while moderate length conflict durations may not have either of these qualities.

### 3.2.5 Survival Support Vector Machine

The Survival SVM method is primarily aimed at predictive performance by optimizing the concordance index (c-index) of a candidate model at the expense of the interpretability of previously implemented methods. Plotting survival curves by strata is not an accessible feature in Survival SVM software. Performance results for this method are provided in Section 3.2.6. It is important to note that there are many options and parameters that can be optimized for various datasets. In this paper a single SVM method is implemented.

Of the three approaches available to model survival data with SVM's, the regression approach is implemented. The kernel type used to fit the model was defined as an additive kernel. In some cases it has been observed that the hybrid of regression and ranking improves the predictive performance [16]. For this paper, the purpose is simply to implement one of the survival SVM approaches as a comparison to the other survival methods.

### 3.2.6 Concordance Comparison of Survival Methods

The predictive performance capabilities of the candidate survival model can be compared using concordance index. The concordance index (c-index) is defined as the ratio of matching pairwise comparisons of observations from all permissible pairs of observation defined by a standardized criterion that addresses issues of ties and censored observations [13]. The four models depicted in Fig. 3.12 consists of the baseline Cox model with all generalized predictor variables, a RSF model with the same variables, the LASSO selected Cox Model, and lastly an example of a survival support vector machine model.

There is compelling evidence that the SVM model outperforms both Cox Models and

the RSF model performed similarly to both Cox Models. The LASSO Cox model has a lower concordance error with improved confidence bands for the random seed utilized. A random seed is required in coding c-indices from bootstrap samples of the PSED. The performance of the RSF may be hindered marginally by some variables from the original selection that are not important, since growing trees with randomly selected variables will grow trees with weak predictive ability. The main purpose here is to identify that implementing the RSF does not jeopardize predictive accuracy relative to the Cox Model, and in many other datasets has been shown to outperform Cox models [13].

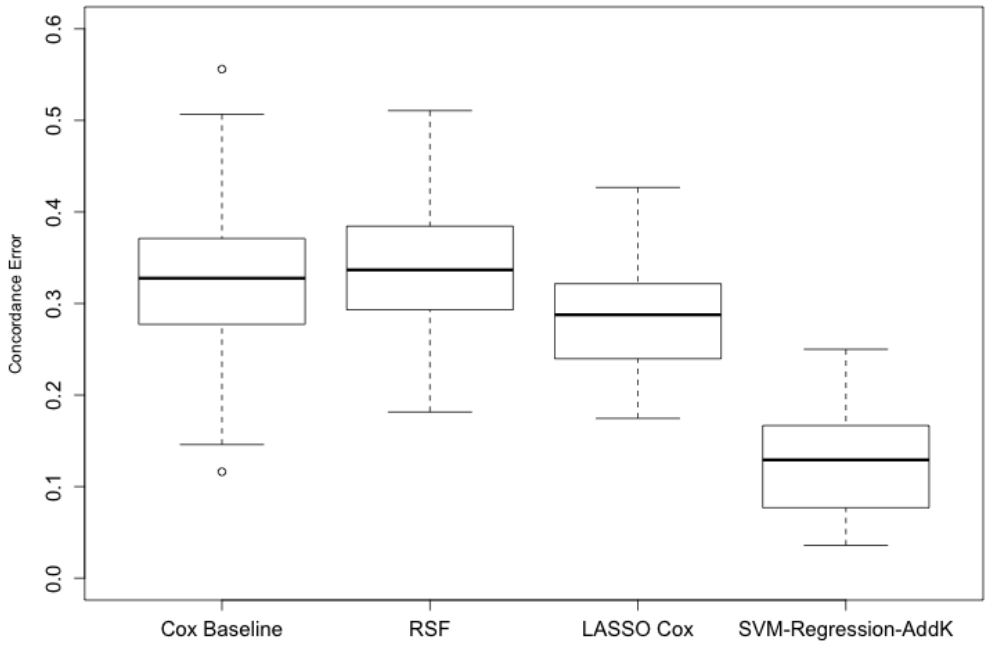


Fig. 3.12: Comparisons of concordance errors rates of candidate survival models. Error rate corresponds to $1 - ConcordanceIndex$. Distribution of error rates correspond to 100 bootstrap samples from the PSED.

## 3.3 Classification Analysis Approach

### 3.3.1 Summary of Classification Methods

This section provides the results for all candidate classification techniques implemented. The performance of all of the methods is summarized in Fig. 3.13. These candidate models have selection algorithms and/or parameters implemented when appropriate in an attempt to increase the interpretability and potentially the predictive accuracy. Classification methods were considered instead of regression methods since many regression methods either do not have right censoring capability or are conventionally used for right censored data. Discretizing the response variable is generally discouraged since some information can be lost to overgeneralizing when categorizing observations. This section will report the results from binary discretization of the response. This is done by considering all right censored observations as "successful" peace agreements and all shorter durations as "unsuccessful" peace agreements. Recall that this can be done for all right censored observations since they are all censored at the end of their respective five-year observational periods.

Table 3.6 reports all candidate models from each method implemented. From an initial inspection of these candidate methods it can be observed that k-Nearest Neighbors with $k = 3$ has the best cross-validated overall error rate. This method classifies exceptionally for true positives which corresponds to peace agreements lasting more than five years having a matching prediction. No other method had this magnitude of sensitivity without jeopardizing specificity by three to ten percent. For strictly predictive purpose this is the optimal classification method.

The other two notable models that have significant interpretability advantages over k-Nearest-Neighbors are the logistic regression and classification tree methods. These methods will have results reported in more detail in Section 3.3.2 and Section 3.3.3. There are some other results in this table that may be initially surprising. These are discussed in more detail in Section 4.1.

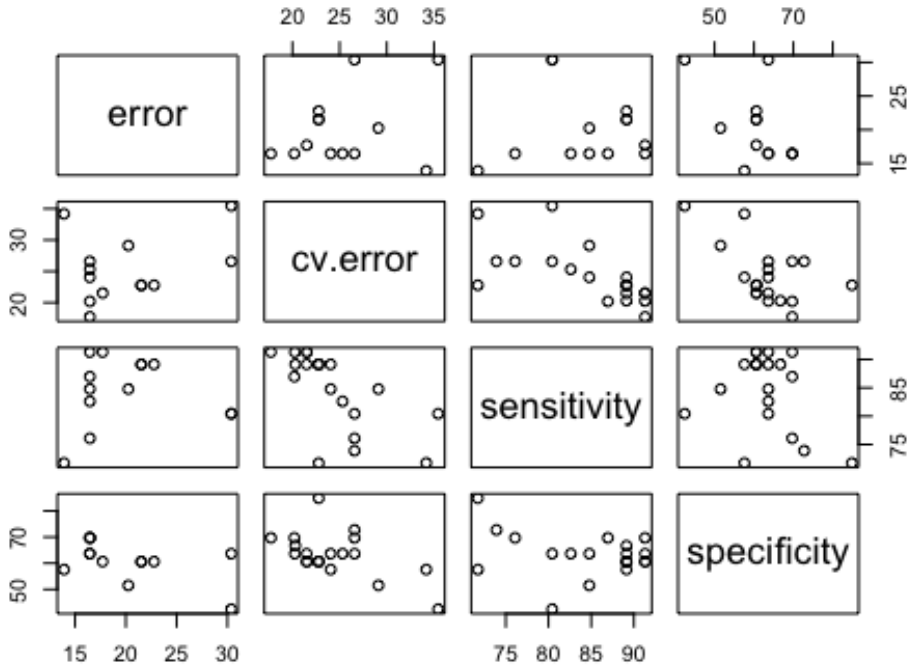


Fig. 3.13: The scatter plot summary of the performance characteristics provides a general understanding of the predictive performance of supervised learning methods for the PSED. For more detailed performance information for each method refer to Table 3.6.

Table 3.6: Summary of Classification Method Performance for Binary Discretized Response. Results correspond to the generalized set of power-sharing variables. Some methods in SAS and R programs report cross validated error rate by default, such as the logistic procedure in SAS. In such cases, no re- substitution error rate is provided. *Mediated Variables and their interaction held fixed in the model during the variable selection procedure.

| Method | Specifications | Error | CV Error | Sensitivity | Specificity |
|---|---|---|---|---|---|
| LDA | Var=14 | 16.46 | 24.05 | 84.78 | 63.64 |
| Stepwise LDA 1 | Var=7 $\alpha = 0.05$* | 22.78 | 22.78 | 89.13 | 60.61 |
| Stepwise LDA 2 | Var=4 $\alpha = 0.05$ | 21.52 | 22.78 | 89.13 | 60.61 |
| Stepwise LDA 2 | Var=7 $\alpha = 0.10$ | 17.72 | 21.52 | 91.30 | 60.61 |
| QDA | Var=14 | 13.92 | 34.18 | 71.74 | 57.58 |
| Stepwise QDA 1 | Var=7 $\alpha = 0.05$ * | 20.25 | 29.11 | 84.78 | 51.52 |
| Stepwise QDA 2 | Var=4 $\alpha = 0.05$ | 21.52 | 22.78 | 89.13 | 60.61 |
| Stepwise kNN | k=3 Var =7 $\alpha = 0.10$ | 16.46 | 17.72 | 91.30 | 69.70 |
| Stepwise Logistic Reg. | Var=7 c=0.56 $\alpha = 0.10$ * | . | 21.50 | 89.13 | 63.64 |
| Stepwise Logistic Reg. | Var=4 c=0.50 $\alpha = 0.10$ | . | 21.50 | 91.30 | 60.61 |
| Backward Logistic Reg. | Var=6 c=0.52 $\alpha = 0.10$ | . | 20.30 | 89.13 | 66.70 |
| GLMM with Blocking | Var=14 c=0.50 | 16.46 | 26.58 | 76.09 | 69.70 |
| GLMM with Blocking | Var=7 c=0.50 | 16.46 | 25.31 | 82.61 | 63.64 |
| GLMM with Blocking | Var=4 c=0.50 | 16.46 | 20.20 | 86.96 | 69.70 |
| Pruned Decision Tree | n=5 | . | 22.78 | 71.73 | 84.85 |
| Random Forest | default | 30.38 | 35.44 | 80.43 | 42.42 |
| Random Forest | mtry=9 ntree=500 | 30.38 | 26.58 | 80.43 | 63.64 |
| Tuned SVM | Var=6 Cost=100 $\gamma = 0.001$ | . | 20.25 | 91.30 | 63.64 |
| Default Boosted Trees | | . | 26.58 | 73.91 | 72.73 |
| Tuned Boosted Trees | | . | 24.05 | 89.13 | 57.58 |

### 3.3.2 Decision Trees

Of the several models implemented in Section 3.3.1, the pruned decision tree (from Table 3.6) has comparable predictive accuracy to many of the methods and provides a highly interpretable model. The selected decision tree was determined using a cp-index plot. In R this plot implements cross-validation on the trees as they are grown. This causes notable variation in the shape of the plot. Using an inspection across several plots it was identified that 3, 5, or 11 trees minimized the tree error. Fig. 3.14 shows one such plot. Using the 1 SE rule to prune for added interpretability it would be plausible to utilize a 3 or 5 node tree. The five node tree was used for demonstrative purposes. Since the cp-plot was extremely unstable, an 11 node tree or larger could be extremely unstable as well as less interpretable.

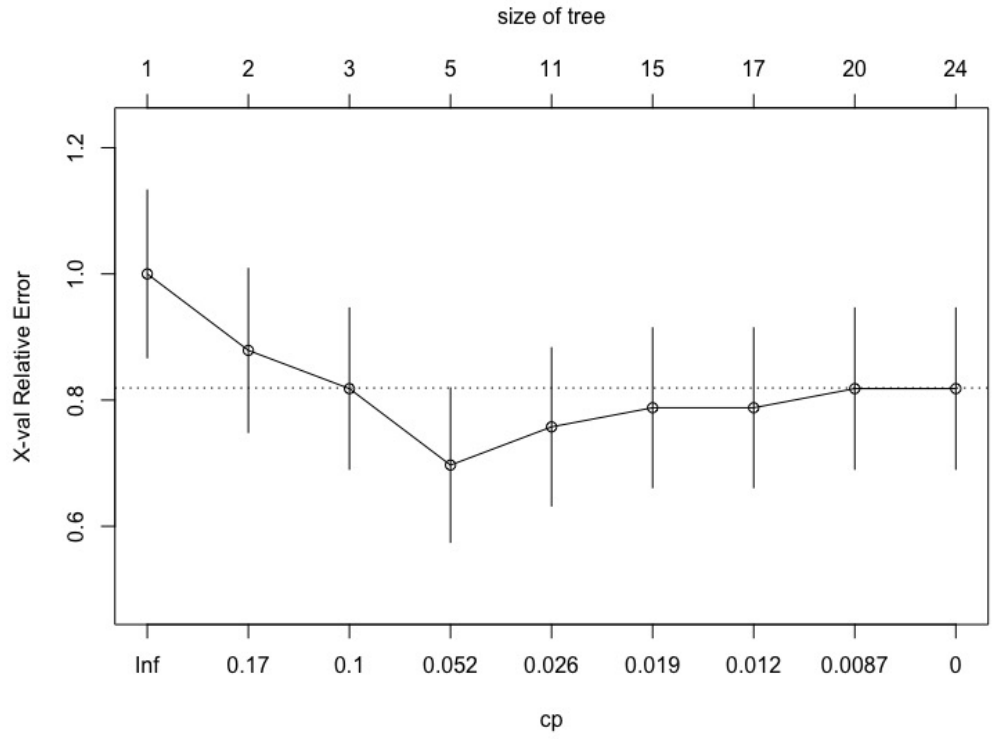


Fig. 3.14: Plot of relative error with respect to the cp-index. Depending on the random seed, 3, 5, or 11 terminal nodes minimized the error. These are all reasonable candidate tree models. The 5 node tree is chosen for this seed since it will provide an interpretably small model with minimized prediction error

Fig. 3.15 shows the pruned 5 node decision tree. The primary split that minimized the Gini Index was across territorial power-sharing. In this model, the tree identifies/predicts that 86 percent of peace agreements that included territorial power-sharing (which comprised 37 percent of all peace agreements) survived longer than 5 years. The remaining node is not as pure, and it is split on the second step of the tree across military power-sharing, and then the final split is chosen across conflict duration. For peace agreements that do include military and territorial sharing, only 25 percent last 5 years or longer. For peace agreements without territorial power-sharing, but include military power-sharing and have a conflict that lasted longer than 986 days, 89 percent lasted longer than 5 years. The the 986 day cutoff (calculated by maximizing daughter node purity) may not lend itself to exact interpretability.

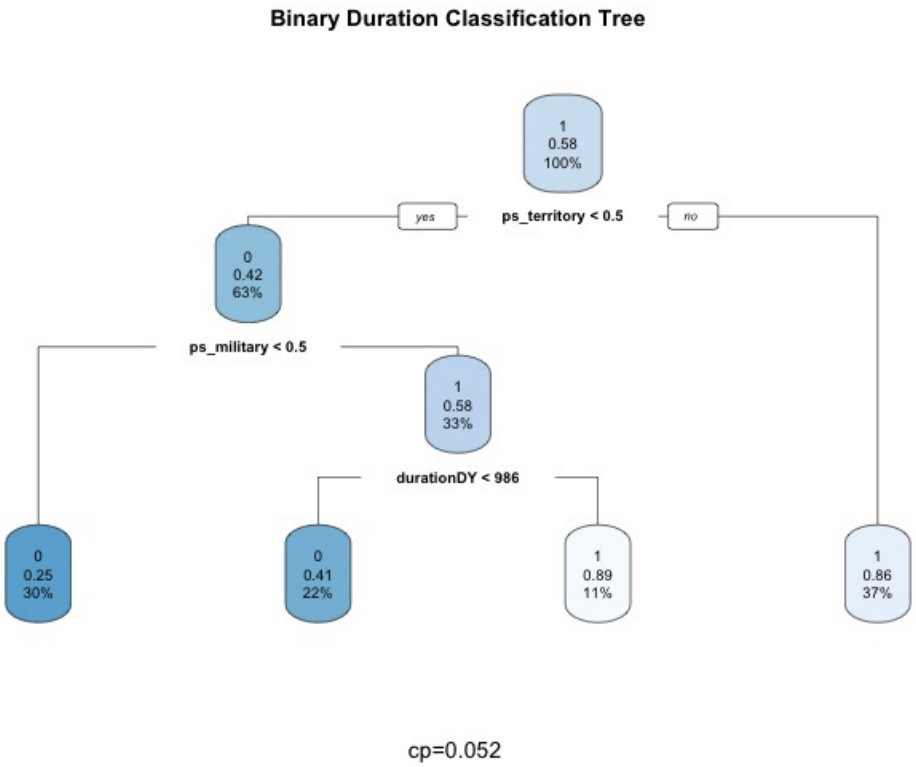


Fig. 3.15: Decision Tree with 5 terminal nodes (cp-index 0.052). Color scale identifies node purity where darker blues correspond to more zero responses (failed peace agreements) and whiter shades correspond to more one responses (successful peace agreements).

Variable importance is also calculated and is displayed in Table 3.7 for the 6 most important variables in the model by minimization of the Gini Index. It can be identified that the 3 variables included in the 5 node tree are determined to be notably more important than the remaining variables. As will be discussed in Section 4.1.4, this is one of the most interpretable models and the predictive accuracies are competitive with most other methods, but there are some important considerations when using Decision Trees.

Table 3.7: Variable importance for 5 node decision tree. Six highest importance variables shown.

| PA Criterion | Importance |
|---|---|
| ps territory | 7.1738 |
| durationDY | 4.5329 |
| ps military | 2.6677 |
| ppsPROM IMP | 1.6712 |
| mpsPROM IMP | 1.5958 |
| intDY | 1.4846 |

### 3.3.3 GLMM: Logistic Regression with Random Blocking Factor

Two approaches were implemented to examine the effectiveness of performing a logistic regression. A standard stepwise regression was performed using Proc Logistic in SAS, and subsequently a generalized linear model was performed in the lme4 package [22] using the glmer function under the assumption of a binomial distribution with a logarithmic link function.

In the SAS logistic regression procedure, cross-validated error rates are reported as a default setting, so resubstituted error rates are not provided in Table 3.6. The backwards eliminated logistic regression had the lowest error rate and a competitive model that falls between the k-Nearest Neighbor and Decision Tree methods with respect to overall error rate. Refer to Table 3.8 for a summary of the backwards selection parameter estimates. For an ROC plot of variable selection steps, refer to Appendix A.4.

Table 3.8: Backwards selected logistic regression maximum likelihood estimates with AUC=0.866. Variable selection using $\alpha = 0.10$ significance level. A logarithmic transformation was performed on durationDY since it improved the model fit marginally.

| PA Criterion | Est. | Std. Error | P>Wald ChiSq | Point Est. | 0.95 Wald CI | |
|---|---|---|---|---|---|---|
| ps territory | -3.2404 | 0.8538 | 0.0001 | 0.039 | 0.007 | 0.209 |
| LNdurationDY | -0.9130 | 0.319 | 0.0042 | 0.4 | 0.215 | 0.750 |
| ps military | -1.9249 | 0.6943 | 0.0056 | 0.146 | 0.037 | 0.569 |
| multireb | 2.1072 | 0.978 | 0.0311 | 8.225 | 1.211 | 55.887 |

In the generalized linear mixed model (GLMM), clustering can be accounted for by accounting for the regional blocking factor that was discussed while implementing hierarchal clustering. Recall that these clusters are not restructured to reflect the peace agreement criterion clustering, but rather they have been grouped generally into 6 international regions. (This will not fully account for the clustered structure, but will do so in an interpretable manner.) A GLMM model may prove to be an effective method in a data set that is almost exclusively categorical and GLMMs are a frequently used class of models for categorical data analysis.

Parameter estimates were constructed in R using the glmmer() function. A generalized linear mixed model needs to be implemented since both fixed peace agreement characteristics and random regional block effects are included in the model. For this method two selections of variables are used to construct a logistic regression: the strictly backwards eliminated subset of variables provided in Table 3.9 (as similarly shown in the previous logistic regression model), and the backwards selected subset where the mediation terms are held fixed in the model provided in Table 3.10. This is done since the mediation variables in many of the classification methods were insignificant (some of them only marginally insignificant). If these are important variables to consider with regards to a specific research question, then keeping them in the model may be worth considering if only to compare to some of the survival methods.

Table 3.9: GLM with backwards selected logistic regression maximum likelihood estimates in the glmer() function in R.

| PA Criterion | Estimate | Std. Error | P>Wald ChiSq |
|---|---|---|---|
| ps territory | -3.2408 | 0.8539 | 0.0001 |
| LNdurationDY | -0.9131 | 0.3188 | 0.0042 |
| ps military | -1.9251 | 0.6943 | 0.0056 |
| multireb | 2.1074 | 0.9777 | 0.0311 |

Table 3.10: Generalized linear model with a binomial distribution and a log link function. Variables selected by backwards selection with mediation variables fixed in the model. Clustering by region was accounted for by including region as a random blocking effect.

| PA Criterion | Estimate | Std. Error | P>Wald ChiSq |
|---|---|---|---|
| med design | 0.427 | 2.672 | 0.7680 |
| med imp | 0.974 | 1.089 | 0.3713 |
| inter | -2.322 | 1.866 | 0.2135 |
| LNdurationDY | -0.9861 | 0.346 | 0.0043 |
| multireb | 2.426 | 1.071 | 0.0235 |
| ps military | -1.879 | 0.725 | 0.0095 |
| ps territory | -3.10 | 0.863 | 0.0003 |

# Chapter 4

# Discussion

The purpose of this chapter is to discuss in more detail the advantages and disadvantages of the methods implemented in relation to the PSED and to identify the optimal candidate methods for various research questions. A critical consideration in the selection of methods that will be discussed in multiple instances is that the PSED is an observational study. Additionally the sample of peace agreements is equal to the overall population. Because of this, choosing the model with the best predictive accuracy is too superficial. In most circumstances it is invalid to predict on observations outside of the population. In our case, it is not statistically sound to make predictions on observations that don't match the criterion of the population that was considered. Recall that all civil war peace agreements were gathered from the post-Cold War era from 1989-2006. This year range may have been chosen for some reason beyond convenience which would mean that it may not be statistically sound to predict on peace agreements outside of this time period.

Because of this unique circumstance, another important consideration may be to examine peace agreement factors that were critical during this era. A retrospective examination for educational or historical purposes may be more important than making future predictions. Model and factor interpretability will be important aspects of model selection to address this circumstance.

## 4.1 The Benefits and Drawbacks of Candidate Methods

In this process of evaluating methods it is important to identify that in almost all cases, it is best to prioritize using survival methods for data that is originally provided as right censored time-to-event data. Classification methods were implemented for comparative purposes, and secondarily to attempt to acquire similar conclusions and exceptional

predictive accuracies from a simplified representation of the response.

A perhaps initially surprising result from Table 3.6 is that Random Forest Classification, Boosted Trees, and Support Vector Machines could not outperform or match k-Nearest Neighbors or some of the other simpler methods. This is most likely due to the almost entirely binary nature of the data set. For the tree-based methods, there is only one split for 14 of the 15 predictor variables. This causes the Random Forest model to grow several similar trees and in many cases less meaningful trees. Having low variance across trees inhibits improvements to predictive accuracies. Additionally, this Random Forest model has also been constructed with a large proportion of variables with negative variable importance. With the repeated random selection of subsets of the predictors at a given node, there will likely be several trees that have poor predictive accuracy, especially if multiple variables with negative variable importance are selected within a given tree.

For the Gradient Boosting Machine the same principle applies that restricts the amount of available splits across a given predictor variable. The weighting of residuals does improve the classification capability of the model in contrast to the Random Forest Model. The improvements did not surpass several of the simpler methods implemented.

It is worth noting that an adaptation of the Random Forest algorithm is available for right-censored data, but evidence has been shown [13] that they do not outperform Cox or SRF models.

For Support Vector Machines, the hyper plane boundary that is optimized can only be divided in a binary n-space in a limited amount of ways. With only one continuous predictor variable in the PSED, the hyperplane is vastly restricted in its ability to adjust boundary strictness/softness and influence of observations. SVM Classification does similarly to the other optimal classification methods and would be reasonable to consider. However, model interpretability and the complexity of parameter tuning may be a considerable drawback since much simpler methods were able to match or outperform SVM classification.

For all three of these advanced classification methods, it is critical to note that there is growing literature extending these methods to the right-censored survival context. The

SRF model has been heavily emphasized in this paper, and Gradient Boosted Survival Trees and Survival SVMs have available software accommodation as of recent years [15, 16, 23].

### 4.1.1 Hierarchal Clustering

Hierarchal clustering was in many aspects able to construct reasonable clusters for the PSED. Even though not fully implemented in this paper, clustered regional structures were identified and could be used in future analysis. It is important to consider that hierarchal clustering is strictly limited to the covariate profile given. If regional clustering is dependent on some other factors beyond what was collected in the PSED (such as a regional economic repression, famine, natural disaster etc.), the hierarchal clustering algorithm will not be able to detect such types of clustered structure. This explains why the algorithm can frequently place observations in clusters that don't have high interpretability.

### 4.1.2 LASSO Cox Model vs. Survival Random Forests

The LASSO Cox and Survival Random Forest models are both cornerstone methods in the analysis of the PSED, and utilization of this dataset promotes a productive discussion comparing these advanced survival methods. They are both included in this section.

The most apparent difference between these two methods was the remarkable difference in the survival curve confidence limits in Figures 3.6 and 3.10. The SRF model was able to clearly detect a difference in complete mediation from the other strata, whereas the Cox Model was unable to do so. The SRF model identified the interaction between mediated design and mediated implementation as the second most important variable, which suggests that splitting across this interaction maximized survival difference across daughter nodes more effectively than most of the peace agreement criteria. The LASSO Cox procedure also identified that the mediation interaction was a highly significant predictor variable in Table 3.5; the coefficient estimate indicates that a one unit increase in the interaction (from 0 to 1) decrease the hazard by a factor of 0.026, but the survival curves lacked the statistical significance across the other strata to suggest that they were different. This is likely because the within strata sample size is relatively small for the "none" and "both" strata (N = 40

and N = 29, respectively) and extremely small for the off-diagonal strata (N = 3 and N = 7).

The bootstrap sampling of the PSED allows for some unique ways to handle the smaller within strata sample size. Observations can gain more representation by sampling without replacement, but more importantly they are used in hundreds to thousands of survival trees with uniquely bootstrapped sampled portions of the PSED. In the case of an observational study like the PSED, sample size is not controllable and there is not a way to improve the confidence limits in a subsequent study as would be the case with other disciplines.

The parameter estimates of the LASSO Cox Model (Table 3.4) are one of the more compelling arguments to continue using Cox regression techniques. Parameter estimates allow for a convenient numerically compact interpretation of the effect of a single covariate. In the LASSO selected model, almost all of the parameter estimates are statistically significant, which validates the use of these coefficients for interpretative purposes. For many theoretical strata that could be constructed for the PSED for other research questions beyond mediation criterion effect, the strata sample sizes are better balanced and allow for strata to be determined to have statistically significant differences in predicted survival curves. In these cases, it would be reasonable to argue that the LASSO Cox Model has the benefits of parameter estimate interpretation without the expense of statistical significance for strata-based research questions for smaller datasets.

The RSF procedure can, in some aspects less conveniently, acquire the same information as parameter estimates through partial dependence plots (as in Fig. 3.11). A hazard ratio less than 1 would correspond to a decrease in predicted mortality count for an increase in a value of the predictor variable. As an example, in Table 3.4, territorial power-sharing has a hazard ratio of 0.172 (which is calculated by exponentiating the adjusted parameter estimate). This identifies that for a 1-unit increase in territorial power-sharing (from zero to one) with all other predictors held constant, the hazard of observing a failed peace agreement decreases by roughly 82.8 percent. In the RSF partial dependence plot for territorial power-sharing (Fig. 3.11), it can be efficiently identified that a one-unit increase

corresponds to a significantly different predicted agreement mortality rate. In the context of the PSED, mortality rate would be considered synonymous to the peace agreement failure rate. Attempting to directly connect this result to the Cox parameter estimate is not necessarily as meaningful or worthwhile as simply identifying that the conclusions match.

In the context of the PSED, using the partial dependence plot (via the RSF procedure as in Fig. 3.11) is contextually meaningful. Acquiring parameter estimates (as in Table 3.4) allows for a precisely predicted quantification of the effect of a predictor. In the context of a non-repeatable observational study such as a post-Cold War era study of civil war duration, these parameter estimates would likely not be implemented in the way that they would in a randomized controlled designed (RCD) experiment. Without assumptions of randomized assignment of treatment, or rather randomized delegation of peace agreement criterion, the interpretability of parameter estimates are jeopardized, and the inferences of the prediction cannot be extended beyond the sampling population. Since all known civil war peace agreements in the post-Cold War Era defined by the years 1989-2006 are included in the study our only assessment of prediction can be performed back on the same peace agreements via test and training datasets or by k-fold cross-validation. The partial dependence plots allow for a quick screening of trends without the need for exact (and arbitrarily precise) estimates of trends across a predictor variable. Partial dependence plots provide a more contextually relevant way of observing how various peace agreement factors influenced the post civil war peace duration.

On a separate note, it could be argued that the assumptions made in the Cox Model are too restrictive in some applications. Generally some degree of violation of the proportional hazard assumption should not disqualify candidate Cox models. Other important considerations are nonlinearity of the predictor variables and the effect of influential observations. Relevant plots are included in Appendix A.3. If severe nonlinearity of predictors exists or if concerning influential observations remain in a Cox Model, the interpretation of the parameter estimates is jeopardized. This can occur if evidence of collinearity is detected among predictors. In an RSF model, no such assumptions are made and no variable

selection procedures are required.

### 4.1.3 Survival SVMs

Fig. 3.12 identified that the Survival SVM using the regression approach to account for right censoring with an additive kernel outperformed the SRF and Cox Models by a significant margin. Other Survival SVM approaches could have been compared and some of them may have had improved prediction accuracies over the approach that was implemented. As an example, it has been shown that generally the hybrid method outperforms the regression approach [16]. For the purpose of this paper, however, it is compelling the Survival SVM should be the primary choice for predictive accuracy.

Drawbacks of this method are considerable with regards to the application of the SVM to the PSED as well as many political science datasets. Interpretability of the model is of considerable importance, and even though SVM theory provides convincing justification for the implementation of this method, visual interpretability of the model is limited to prediction accuracy and a plot of survival curves (which is not provided in this paper). Additionally, predictive accuracy is not necessarily as important for observational studies with a sample size equal to the population. The SRF and the LASSO Cox Models provide important information that is not as accessible in the SVM algorithm.

### 4.1.4 k-Nearest Neighbors

The candidate kNN model using stepwise variable selection achieved the lowest overall error rate (Table 3.6). This model performed exceptionally for detecting true positives with a sensitivity reading of 91.30 percent. If predicting successful peace agreements is the priority, then few methods are comparable. The detection rate for true negative is slightly above average in comparison to the other methods implemented. Given that the PSED is almost strictly binary and clustering structure exists among peace agreements, it is reasonable to understand how voting predictions of agreements off of the three nearest neighbors (three most similar agreements by their criteria) could provide a meaningful and accurate prediction of a peace agreement's duration.

Beyond predictive capabilities, a deficiency of this relatively simple algorithm is that there is not as much information to be gained with regards to model interpretability. The stepwise procedure for variable selection allows some notion of important variables in the model through the assessment of the significance of a predictor effect. It is important to distinguish that more statistically significant variables are directly correlated with the importance measurements found in tree methods, which involve an optimization of survival difference, Gini Index, or some other metric. The variables that remained in the stepwise procedure have strong similarities to the LASSO Cox regression. For meaningful visualizations and a more extensive understanding of the PSED, other methods are required.

### 4.1.5 Decision Tree

Using a pruned decision tree has the benefit of a comparable, though not exceptional, predictive capabilities while creating a highly interpretable model. In the PSED, classification trees retained the highest achieved detection rate for true negatives of 84.85 percent (Table 3.6). The decision tree allowed for an understanding of which variables are the most important in maximizing daughter node purity, which corresponds to improved classification of peace agreements. The visualization of the classification tree (Fig. 3.15) is advantageous since it is an efficient way to understand proportions of observations with various criteria and what proportions were identified to have failed or lasting peace agreements.

Some notable disadvantages of decision trees are that they are relatively unstable and can have relatively significant changes in predictions for small perturbations in the data. Since trees are a greedy algorithm, growing only one tree may inhibit another meaningful and competitive tree from taking precedence. Note that unlike the SRF model, the mediation variables in Table 3.7 were not identified as important variables. This may be evidence to suggest that discretizing the response in the context of the PSED is not an appropriate course of action.

### 4.1.6 Logistic Regression

The logistic regression methods show evidence of comparable and improved error rates

compared to the other various methods (Table 3.6). Similar to the Cox Regression procedure in the survival analysis portion of the paper (Section 3.2), logistic regression yields parameter estimates in a relatively interpretable model. The logistic regression construct using a generalized linear model function allows for clustering to be accounted for by blocking peace agreements by the regional variable discussed previously.

There is evidence that the logistic regression approach combines competitive prediction rates and interpretability. However, unlike the LASSO Cox model, the mediation terms and interaction in the model have no evidence of having a significant effect on the peace agreement's success or failure. This is further confirmation that using binary classification is not a realistic simplification of the PSED since there is evidence that the discretization of the response resulted in a loss of information that jeopardized the identification of a significant effect for some predictors.

## 4.2 Research Questions and Optimal Methods

When considering what research question can be examined for the PSED (and which of the considered methods would be most appropriate for those questions), it is critical to revisit that the dataset is collected as an observational study. This limits the power of the conclusions, to only the identification of associative relationships, rather than causative relationships between predictors and the response. All of these questions can only be interpreted in this context. This context is a frequent occurrence in disciplines such as Political Science where randomized, controlled, and ideally double blind experiments are not possible to arrange. Associations may be identified and used to argue the justification of future policy, awareness, and/or understanding, with the acknowledgement that a claim of causation is not achievable by the limitations of the data.

### 4.2.1 What peace agreement factors are the most important in determining the duration of a peace agreement?

The Survival Random Forest (RSF) procedure provides the most meaningful identification of variable's importance (Fig. 3.9). Growing many weak learning, such as survival

trees, the stability of the model is improved at the expense of interpretability. This method was exclusive in its ability to identify the mediation interaction to be the second most important variable in the PSED. In the Cox approach the p-value significance (Table 3.4) does not have a direct or meaningful connection to the actual importance of a variable in making accurate predictions.

### 4.2.2 What is the individual effect of a peace agreement factor? Is this effect significant?

The LASSO Cox model allows for a quantifiable and readily interpretable measure of the effect of an individual peace agreement factor on the hazard rate of observing a failure (Table 3.4). An appropriate discussion presents itself in the context of the PSED and other observational studies that use census data, where quantifying an arbitrarily exact parameter estimate doesn't necessarily provide further meaning beyond identifying a significant predictor-response relationship.

### 4.2.3 What categories or strata of peace agreements have the highest probability of survival?

The RSF procedure was the only method implemented that was able to detect a significant difference between mediation strata (Fig. 3.10). For highly unbalanced strata with small counts in some strata, there is evidence that the RSF is the optimal choice. In many cases, however, a Cox regression procedure would likely be sufficient. Note that even for the "None" and "Both" mediation strata which had between 20 and 40 peace agreements, there was still notable overlap in the confidence intervals where the RSF model identified an explicit and notably larger difference in strata survival probability (Fig. 3.6 and 3.10). The Survival SVM does not provide accessible visualization, so this method is not currently as effective at evaluating strata.

### 4.2.4 How effectively can a peace agreement's response (failure or no observed failure) be predicted?

For survival methods, the SVM model outperformed all other models (Fig. 3.12) without any elimination of variables for candidate survival tree node splits and without placing any assumptions on the model. For strict predictive accuracy, the SVM is the optimal choice.

For classification methods, using k-Nearest Neighbors, logistic regression, and classification trees all provided some degree of meaningful interpretation paired with competitive predictive accuracies (Table 3.6). The choice of classification methods is highly dependent on unique attributes of a given dataset. If classification is being considered for another dataset, the majority, if not all, of these methods should be implemented.

### 4.2.5 What types of clustering exist in the data?

An unsupervised learning approach, specifically agglomerative clustering, is a valuable tool for understanding the grouping trends in the data. Using the heat map visualizations allow for the assessment of clusters by the predictor variable qualities they possess. Clustering in some applications is more obvious than others. In the context of the PSED there are many ways to cluster the peace agreements. These include grouping by Dyadic party, country, region etc. With agglomerative clusters these grouping characteristics could each be verified to be related to clustering by the civil war peace agreement factors.

### 4.2.6 Which peace agreement factors can classify or predict peace agreement response best?

Tree-based methods and most classification methods can adequately address this question. Variable importance is the most interpretable for assessing the effect by individual predictors since splitting criterion involves some type of optimization of purity or survival difference. Examining ROC curves in a stepwise selection procedure (as in Appendix A.4) can assist in assessing the fit of a subset of variables without the variable that is considered for elimination.

### 4.2.7 Can certain predictors be easily interpreted or screened to better model interpretability or to understand their individual classification effect?

The LASSO Cox model is a considerably useful tool for screening out collinear variables and simplifying the model for interpretability. In our case half of the variables were removed and the c-index was not jeopardized.

A classification tree is an excellent tool to gain further understanding of the variables that are utilized in a pruned tree. This is also relatively accessible in the SRF model.

## 4.3 Conclusion

### 4.3.1 What is the role of Statistical Learning in Political Science?

Political Science as a discipline in many aspects is worlds apart from clinical trials, a field where statistics comfortably resides. "Treatment" can not always be randomly or blindly assigned, and often political scientists study research questions observationally rather than experimentally.

In Chen's thesis [4], it is mentioned that "we still know little about whether, how, when, and why peace agreements can produce their expected effects." In the post-Cold War era, an exponential rise in peace agreements occurred in comparison to the entire Cold War era. A surprisingly high rate of failure occurred across these peace agreements that resulted often in escalated violence. The question arose: "Why do some civil war settlements break down within months whereas others produce lasting peace?"

In the process of addressing this question statistically, it has been reiterated that casual inference will not be attainable. A political scientist can study a dataset such as this with the intent to identify significant predictor-response relationships that identify association rather than causation.

Statistical learning methods provide many benefits when approaching this type of data, including the following:

1. There are many accessible and interpretable methods available. Very often a single

method will not sufficiently meet the demands of a research question. Having a wide array of methods in this paper allowed for multiple candidate models to understand the dataset and find that the SRF model handled the smaller strata sample sizes better than the Cox models which allowed for more significant detection of differences across mediation strata.

2. Model performance for a given dataset can be evaluated in relation to other methods more effectively when several methods are applied. It is often difficult to gain a sense of how well a single model predicts for a given dataset without other methods being implemented for comparison.

3. The dataset can be understood through studying clustered structures, model performance/predictive accuracies, variable importance, and partial dependence.

With these benefits in mind, the Survival Random Forest Algorithm should be an essential tool for the political scientist that studies time-to-event data. Using advanced extensions of the Cox family of methods in partnership with more recent survival analysis extensions of statistical learning methods provides a sound foundation for consolidating/confirming answers to specialized research questions.

### 4.3.2 Future Work

Other survival methods are available that have not been implemented in this paper that may be worth considering in future analysis of this dataset. The Cox procedure can have variable reduction and regularization performed using the Elastic Net constraints [24]. If a further investigation into Cox parameter estimates suggested that some estimates are not realistic or counter-intuitive, a regularization procedure such as Ridge Regression or the Elastic Net may resolve these issues. Regularization may also resolve some of the issues of parameter estimate divergence that was observed in Fig. 3.3.

Boosted survival models are also worth considering for future implementation. A Boosted Cox Model can be performed using the mBoost, CoxBoost, or glmBoost packages. Additionally, a nonparametric survival gradient boosting machine can be constructed

using the GBMCI package in R.

Implementing this wide array of survival methods on several time-to-event political science datasets should also be considered to demonstrate the performance of such methods across a diverse range of context and unique data characteristics.

The Survival SVM's comparison to the RSF model also has more work that could be done. Plotting survival probability curves by strata in the Survival SVM procedure would allow a comparison of the two methods in their ability to distinguish survival differences across strata. This is not a readily accessible feature in the SVM software.

# References


[1] D. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society*, vol. 34, pp. 187–220, 1972.

[2] J. Box-Steffensmeier and B. Jones, *Event History Modeling: A Guide for Social Scientist*. New York: Cambridge, 2004.

[3] M Ottman and J Vullers, "Power-Sharing Event Dataset (PSED)," Datorium [Online]. Available: `http://dx.doi.org/10.7802/69`.

[4] C. Chong, "Negotiated settlement and the durability of peace: Agreement design, implementation, and mediated civil wars," Master's thesis, Utah State University, Logan, UT, 2015.

[5] K. Bogaerts, A. Komarek, and E. Lesaffre, *Survival Analysis with Interval Censored Data: A Practical Approach with Examples in R and SAS and BUGS*. Boca Raton, Florida: Taylor and Francis Group, 2018.

[6] J. Ward and M. Hook., "A hierarchical grouping procedure applied to a problem of grouping profiles."

[7] O. Maimon and L. Rokach, *Data Mining and Discovery Handbook 2nd Edition*, ch. A survey of Clustering Algorithms.

[8] R. Tibshirani, "Regression shrinkage via the lasso."

[9] R. Tibshirani, "The lasso method for variable slection in the cox model."

[10] R. Peto, "Experimental survival curves for interval censored data," *Journal of Statistical Software*, vol. 39, pp. 86–91, 2011.

[11] Y So and G Johnston and S Kim, "Analyzing Interval Censored Survival Data with SAS Software," "SAS Global Forum", 2010.

[12] D. Finklestein, "A proportional hazards model for interval-censored failure time data," *Biometrics*, vol. 42, pp. 845–854, 1986.

[13] H. Ishwaran, U. K. E. Blackstone, and M. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, pp. 841–860, 2008.

[14] H. Ishwaran and U. Kogalur, "Random survival forest for r," *CRAN Rnews*, vol. 7/2, pp. 25–31, 2007.

[15] C. Fouodo, I. Konig, C. Weihs, A. Ziegler, and M. Wright, "Support vector machines for survival analysis with r," *The R Journal*, vol. 2, pp. 841–860, 2008.

[16] V. VanBelle, K. Pelckmans, S. VanHuffel, and J. Suykens, "Support vector methods for survival analysis: a comparison between ranking and regression approaches," *Artificial Intelligence in Medicine*, vol. 53, 2018.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013.

[18] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *cluster: Cluster Analysis Basics and Extensions*, 2018, r package version 2.0.7-1 — For new features, see the 'Changelog' file (in the package source).

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011 [Online]. Available: `http://www.jstatsoft.org/v39/i05/`.

[20] T. M. Therneau, *A Package for Survival Analysis in S*, 2015, version 2.38 [Online]. Available: `https://CRAN.R-project.org/package=survival`.

[21] J. Ehrlinger, "ggrandomforests: Exploring random forest survival," 2016.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] Y. Chen, Z. Jia, D. Mercola, and X. Xie, "A gradient boosting algorithm for survival analysis via direct optimization of concordance index," *Computational and Mathematical Methods in Medicine*, 2013.

[24] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox?s proportional hazards model via coordinate descent," *Journal of the Royal Statistical Society*, vol. 22, pp. 1–13, 2011.

# Appendices

# Appendix A

# Additional Results

## A.1 Detailed Power Sharing Survival Analysis Results: LASSO Cox

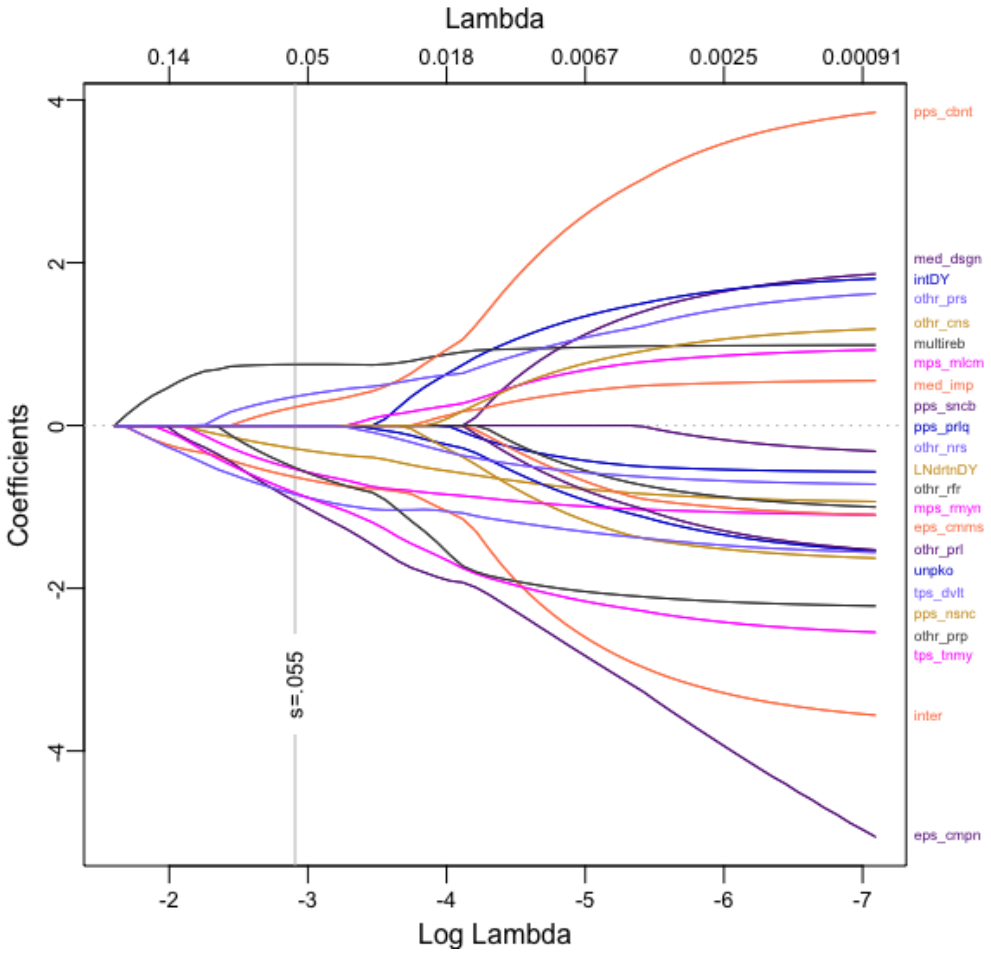


Fig. A.1: LASSO trace plot for detailed power-sharing variables, with all non-power-sharing variables remaining in the model. Evidence of collinearity among some of the predictors. LASSO constraint optimized at s=0.055.

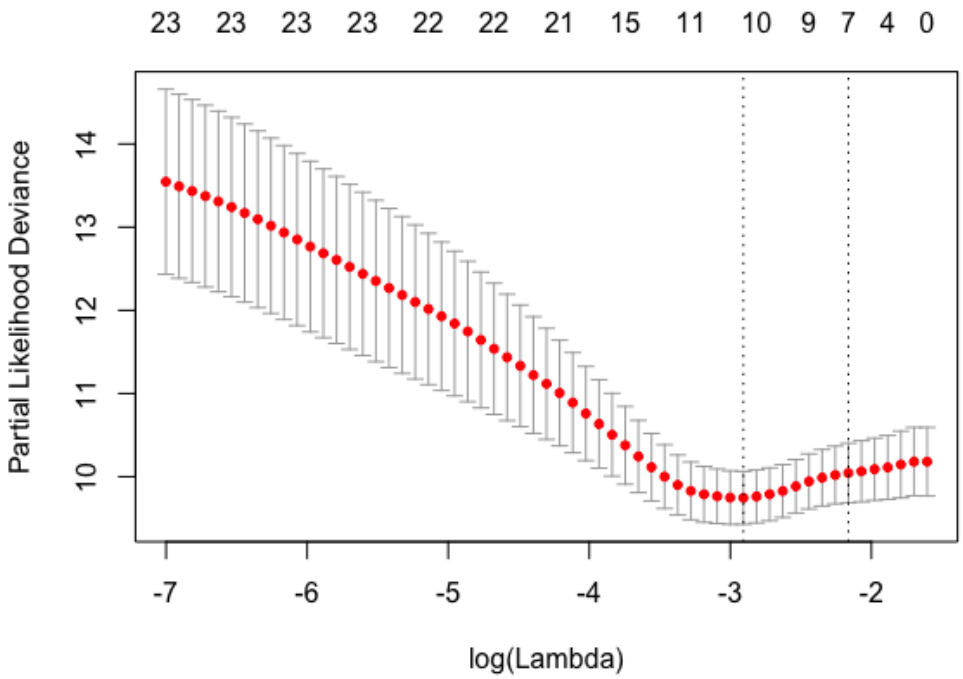


Fig. A.2: k-Fold Cross Validation Plot of LASSO variable reduction. Reduction of variables optimizes model fit at 10 variables.

Table A.1: Coefficients for LASSO Selected Detailed Power Sharing Variables. Note that eps company has a concerning and highly insignificant p-value. In the trace plot, eps company enters the model early and increases linearly without any indication of convergence. This may be evidence that collinearity is not effectively eliminated through the LASSO algorithm. Ridge Regression or Elastic Net may resolve this problem. This variable could also be eliminated through a backwards elimination procedure. Elimination of this variable will not jeopardize model fit as shown by the LASSO k-fold plot.

| PA Criterion | Est. | Hazard Ratio | SE(Est.) | z | p |
|---|---|---|---|---|---|
| med design | 1.26 | 3.51 | 0.814 | 1.54 | 0.123 |
| med imp | 1.13 | 3.09 | 0.608 | 1.86 | 0.063 |
| inter | -3.27 | 0.0382 | 1.13 | -2.89 | 0.004 |
| LNdurationDY | -0.715 | 0.489 | 0.216 | -3.31 | 0.0009 |
| multireb | 0.881 | 2.41 | 0.550 | 1.60 | 0.11 |
| pps cabinet | 0.752 | 2.12 | 0.531 | 1.42 | 0.157 |
| mps armyint | -1.02 | 0.362 | 0.423 | -2.40 | 0.016 |
| eps company | -18.10 | 1.45e-08 | 6.10e+03 | 0.00 | 0.998 |
| tps devolution | -1.42 | 0.243 | 0.658 | -2.15 | 0.031 |
| tps autonomy | -1.94 | 0.144 | 1.06 | -1.82 | 0.069 |
| other proprep | -1.24 | 0.290 | 1.04 | -1.19 | 0.233 |
| other preselect | 0.611 | 1.84 | 0.467 | 1.31 | 0.191 |

Table A.2: LASSO Selected Cox parameter estimates with eps company removed through 1-step of backwards elimination.

| PA Criterion | Est. | Hazard Ratio | SE(Est.) | z | p |
|---|---|---|---|---|---|
| med design | 1.31 | 3.71 | 0.820 | 1.60 | 0.110 |
| med imp | 1.16 | 3.20 | 0.607 | 1.92 | 0.055 |
| inter | -3.39 | 0.034 | 1.14 | -2.97 | 0.003 |
| LNdurationDY | -0.781 | 0.458 | 0.216 | -3.62 | 0.0002 |
| multireb | 0.900 | 2.46 | 0.540 | 1.67 | 0.095 |
| pps cabinet | 0.610 | 1.84 | 0.526 | 1.16 | 0.246 |
| mps armyint | -1.16 | 0.312 | 0.413 | -2.82 | 0.005 |
| tps devolution | -1.70 | 0.183 | 0.650 | -2.61 | 0.009 |
| tps autonomy | -1.89 | 0.150 | 1.07 | -1.77 | 0.076 |
| other proprep | -1.26 | 0.285 | 1.04 | -1.20 | 0.228 |
| other preselect | 0.641 | 1.90 | 0.478 | 1.34 | 0.179 |

## A.2 Detailed Power Sharing Survival Analysis Results: SRF

The SRF model had an initial variable screening since many variables had no information recorded. Twenty-three variables are included for the analysis. Concordance performance for this method and the LASSO Cox model was not compared between this model and any other models in this paper.

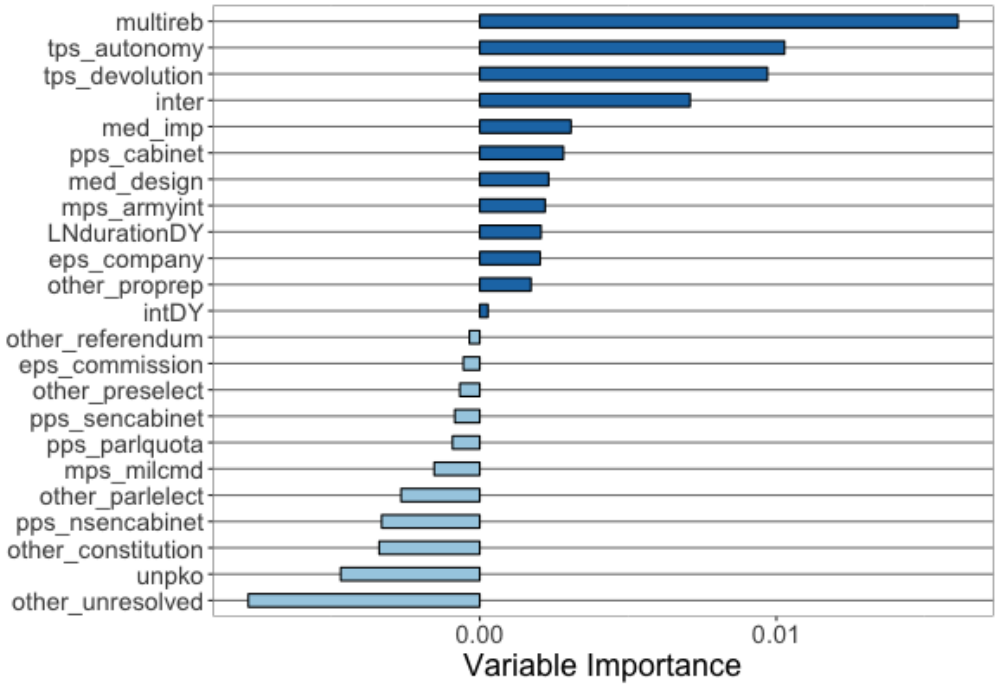


Fig. A.3: SRF Model for the with the detailed power-sharing variables replacing the generalized set. Note that the magnitude of variable importance is much lower in this model than in the SRF presented for the generalized power-sharing variables, but that the mediation interaction term is still one of the most important variables in the model. The territorial power sharing variables show evidence of remaining some of the most important variables in the model. Conservation of Events criterion is still used in this model.

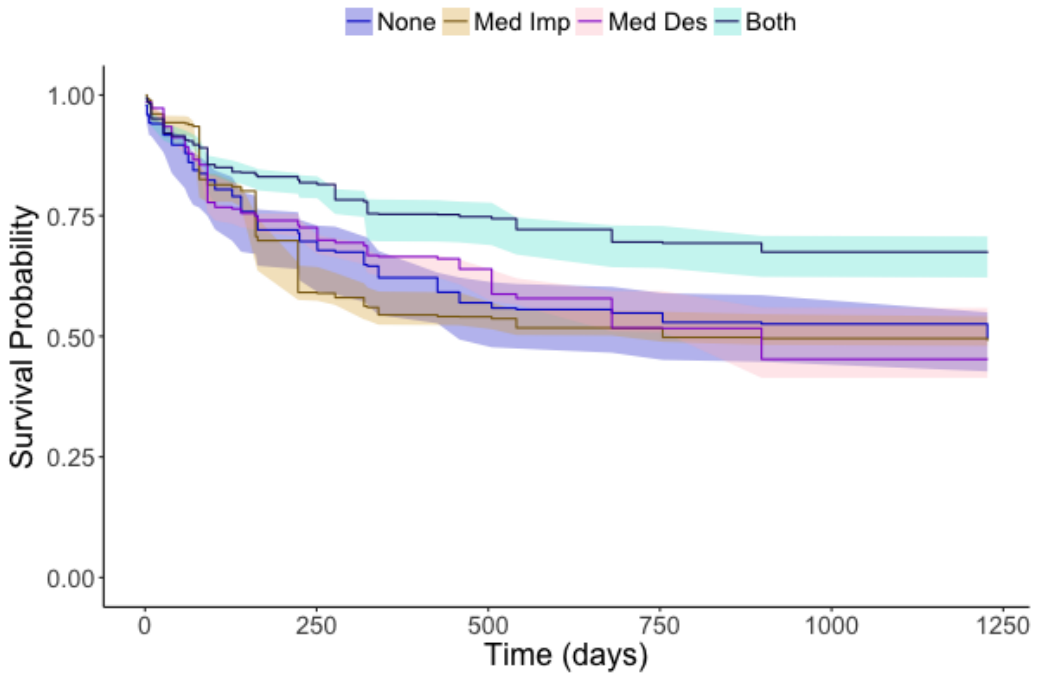


Fig. A.4: SRF Model for the detailed power-sharing variables is still able to detect a significance between full mediation in the peace agreement process from other mediation strata.

### A.3 Cox Model Diagnostics Graphics

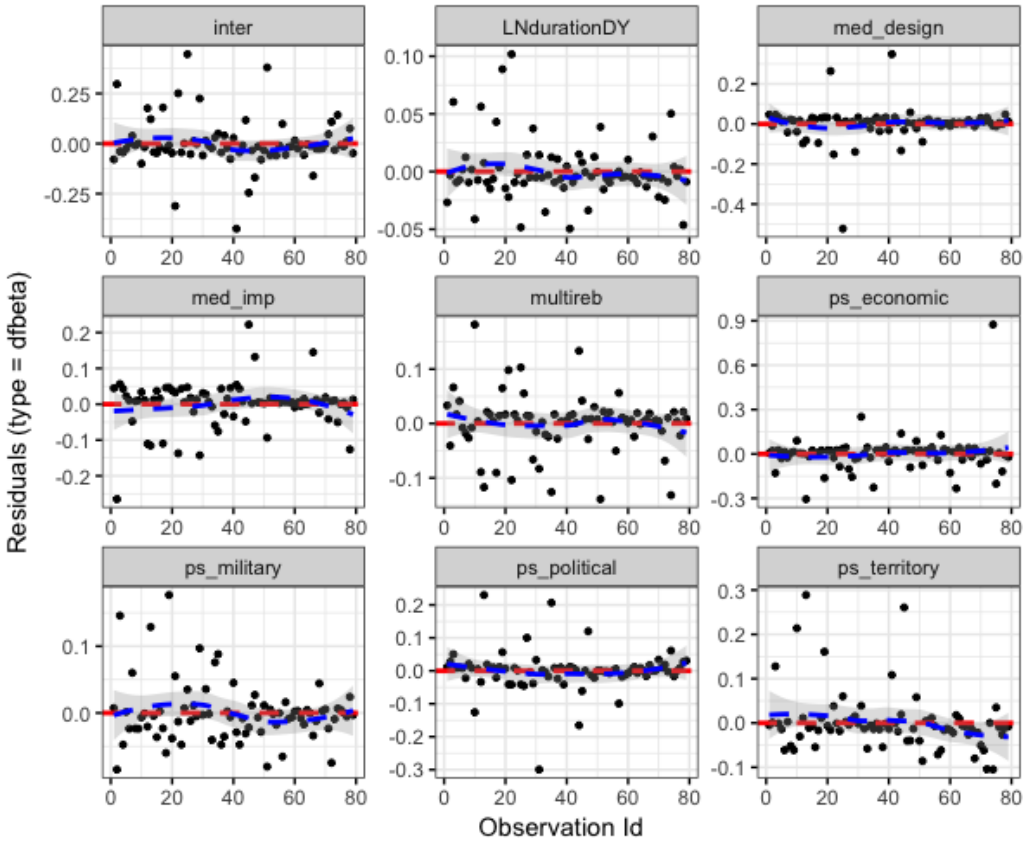


Fig. A.5: Inspection of influential observations by peace agreement factors. No influential points were removed in this paper for any analysis. Influential points may have significant effects on parameter estimates in regression procedures, and the removal of such points should be considered to improve interpretability of the model.

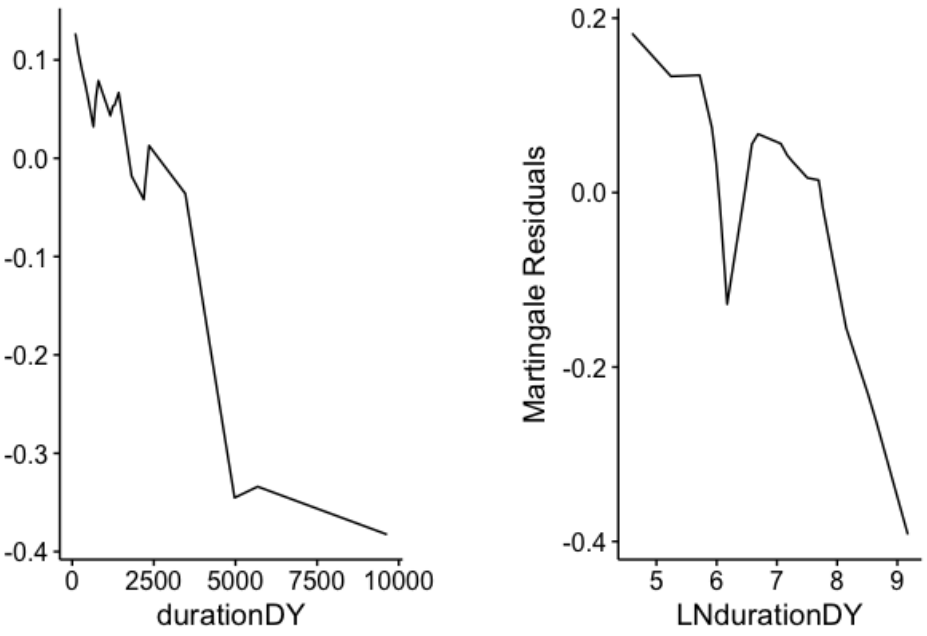


Fig. A.6: Martingale Residuals are plotted against continuous covariates to test for nonlinearity. Conflict Duration (DurationDY or LNDurationDY) is the only continuous covariate in the model. There is weak evidence that nonlinearity exists for this predictor after performing a log transformation. The natural log transformation may have been too strong of a transformation for this Conflict Duration. A square root or cube root transformation, though less frequently implemented, may be the more optimal transformation.

## A.4 Logistic Regression: ROC Curves Examination

Fig. A.7 shows the decrease in model fit by variable eliminated in the backwards selection procedure. There is overall loss in the AUC by +0.07. This may be too large of a loss, and Step 8 or Step 9 may be a better choice. This 4-variable model, however, still retained high predictive accuracy and with high model interpretability and is worth considering as a candidate model.

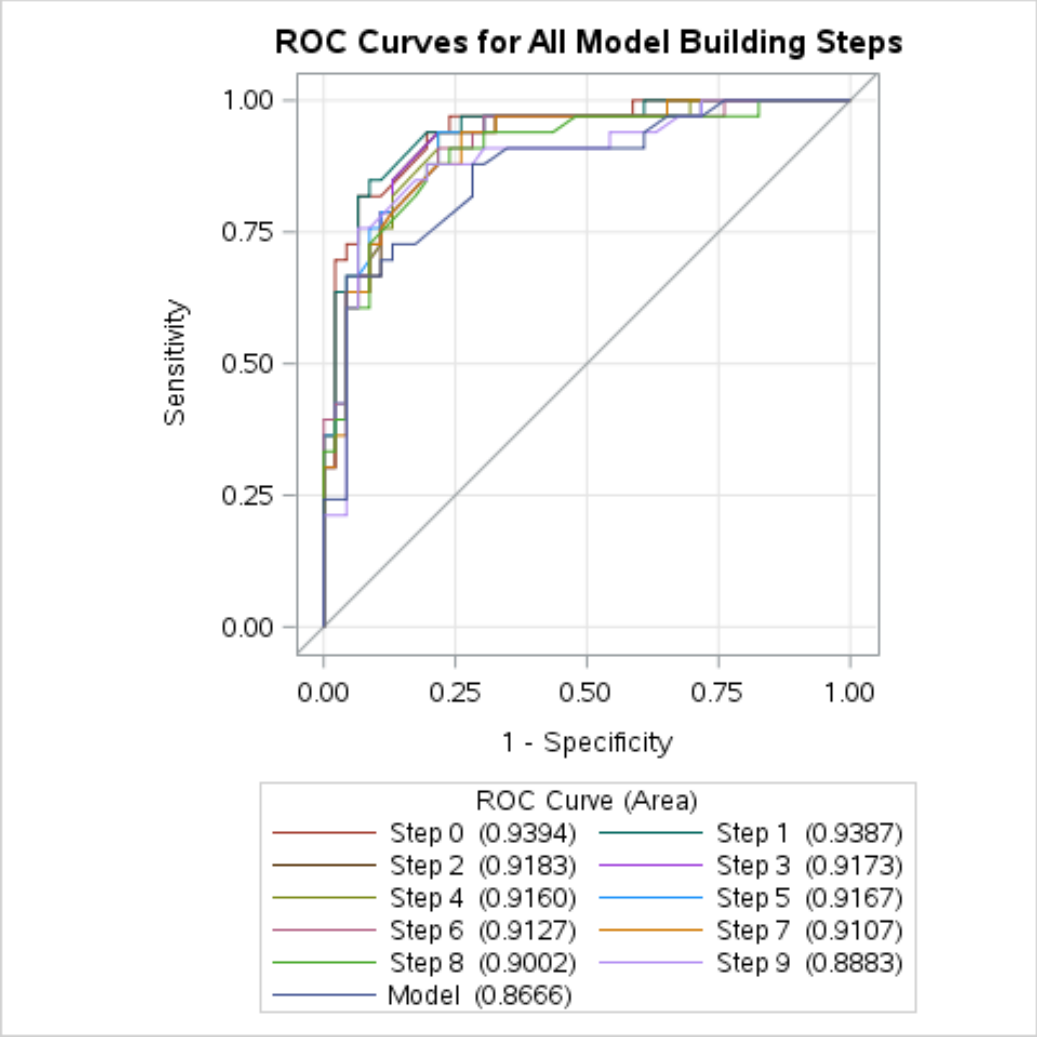


Fig. A.7: ROC curves by step in the Backwards Selected Logistic Regression.

# Appendix B

# R and SAS Code

```
PSED<- read.csv("PSEDTHESIS.csv")
PSED$durationDYStd<-(PSED$durationDY-mean(PSED$durationDY))/sd(PSED$durationDY)
PSED$LNdurationDY<-log(PSED$durationDY)
PSED.clean<-PSED[,-c(19,21,23,25,27,29,31,33,35,37,39,41,43,45,47)]
```

## B.1 Hierarchal Clustering Code

```
library (cluster)
library (purrr)
library (gplots)
library (RColorBrewer)

m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

# function to compute agglomerative clustering coefficient
ac <- function(x) {
  agnes(PSED.clean[ , c(2:5, 7:16,18:33, 40)], method = x)$ac
}
map_dbl(m, ac)

row.names(PSED.clean) = make.names(PSED$location, unique = TRUE)
hc1 <- agnes(PSED.clean[ , c(2:4, 6:16, 40)], method = "ward")
```

```
par(mar = c(2.5, 2.5, 1.5, 3))
plot(as.dendrogram(as.hclust(hc1)),
  nodePar = list(lab.cex = 0.60, pch = NA),
  main    = "Agglomerative Clustering: Ward's Method",
  horiz   = TRUE,
  xlim    = c(12, 0))

par(cex.main = 1.10, cex.axis = 0.85)
heatmap(as.matrix(PSED.clean[ , c(2:4, 6:16, 39)]),
        main   = "Cluster Characterization with Generalized Power-Sharing
        Variables",
        cexCol = 1.10,
        cexRow = 0.95,
        Colv   = NA,
        scale  = "none",
        col    = brewer.pal(9, "BuPu"))

#For other Dendrograms
hc2 <- agnes(PSED.clean[ , c(2:4, 6:16, 18:33, 40)], method = "average")

par(mar = c(2.5, 2.5, 1.5, 3))
plot(as.dendrogram(as.hclust(hc2)),
  nodePar = list(lab.cex = 0.45, pch = NA),
  main    = "Agglomerative Clustering: Average Linkage",
  horiz   = TRUE,
  xlim    = c(12, 0))

hc3 <- agnes(PSED.clean[ , c(2:4, 6:16, 18:33, 40)], method = "single")
```

```
par(mar = c(2.5, 2.5, 1.5, 3))
plot(as.dendrogram(as.hclust(hc3)),
  nodePar  = list(lab.cex = 0.45, pch = NA),
  main     = "Agglomerative Clustering: Single Linkage",
   horiz   = TRUE,
   xlim    =  c(12, 0))

hc4 <- agnes(PSED.clean[ , c(2:4, 6:16,18:33, 40)], method = "complete")

par(mar = c(2.5, 2.5, 1.5, 3))
plot(as.dendrogram(as.hclust(hc4)),
  nodePar  = list(lab.cex = 0.45, pch = NA),
   main    = "Agglomerative Clustering: Complete Linkage",
   horiz   = TRUE,
   xlim    = c(12, 0))
```

### B.2 Survival Analysis Code

```
library(glmnet)
library(survival)
library(ggplot2)
library(usdm)
library(plotmo)

PSED.sub <- PSED[ , c(2:17, 48, 52)]
vif(PSED.sub[ , c(1:3, 4:15)]) #Generalized Set of Vars
vif(PSED.sub[ , c(1:2, 4:7, 17:39)]) #Non-Generalized Set of Vars
vif(PSED.sub[ , c(1:15, 17:22, 30:35)])#Almost All Vars
```

```
cor(PSED.sub[ , c(11,15, 33, 35)]) #portion of tps vars

cox.B <- coxph(Surv(duration, status) ~ med_design + med_imp + inter +
              intDY + LNdurationDY+ unpko+ multireb + ps_political +
              ps_military + ps_economic + ps_territory + ppsPROM_IMP +
              mpsPROM_IMP + epsPROM_IMP + tpsPROM_IMP,
              data = PSED.clean)
cox.B

#Fit LASSO COX PH MODEL
x  <- model.matrix( ~ med_design + med_imp + inter + intDY +
        durationDY + unpko + multireb + ps_political + ps_military +
        ps_economic + ps_territory + ppsPROM_IMP + mpsPROM_IMP +
        epsPROM_IMP + tpsPROM_IMP,
        data=PSED.clean)
y  <- Surv(PSED$duration, PSED$status)
fit      <- glmnet(x, y, family = "cox", alpha = 1)
allnames <- names(coef(fit)[ , ncol(coef(fit))]
                    [order(coef(fit)[ , ncol(coef(fit))], decreasing=TRUE)])
allnames <- setdiff(allnames, allnames[grep("Intercept", allnames)])
cols     <- rep("gray", length(allnames))

plot_glmnet(fit,
  label = TRUE,
  s     = cv.fit$lambda.min,
  col   = c("coral", "darkorchid4", "blue3", "slateblue1",
            "goldenrod3", "gray27",  "magenta")))
```

```
cv.fit <- cv.glmnet(x, y, family = "cox", alpha = 1)

plot(cv.fit)
coef(cv.fit,s = "lambda.min" )

cox.L <- coxph(Surv(duration, status) ~ med_design + med_imp + inter +
                LNdurationDY + multireb + ps_political + ps_military +
                 ps_economic + ps_territory ,
                 data=PSED.clean)
cox.L
coef(cox.L)

ggcoxdiagnostics(cox.L , type = "dfbeta")
ggcoxfunctional(cox.L, data = PSED.clean)
aa.fit <- aareg(Surv(duration,status) ~ med_imp + med_design + inter +
                       LNdurationDY + multireb + ps_military + ps_economic +
                        ps_territory,
                        data = PSED.clean)
aa.fit
res.cox <- cox.zph(cox.L)
autoplot(aa.fit, palette = "Accent") +
  list(ggplot2:: scale_color_manual(values = c("black", "black", "black",
                          "black",  "black", "black", "black", "black", "black")),
       ggplot2:: scale_fill_manual(values = c("chocolate4", "gray27",
                          "darkorange", "yellow", "turquoise1","darkviolet",
                          "magenta", "mediumblue", "plum"))) +
   xlab("Time (days)") +
   ylab("Schoenfield Residuals Beta(t)") +
```

```
   theme(panel.background = element_rect(fill = "white", colour = "black",
                size = 0.5, linetype = "solid"),
                panel.grid.major = element_blank(),
                panel.grid.minor = element_blank())

cox.L.strata <- coxph(Surv(duration, status) ~ strata(med_design) +
                  strata(med_imp) + LNdurationDY + ps_military +
                  ps_territory+ ps_economic +multireb,
                  data=PSED.clean)
cox.fit.LS   <- surv_fit(coxL.strata, data = PSED.clean)

ggsurvplot(cox.fit.LS,
  size     = 0.5,
  data     = PSED.clean,
  conf.int = TRUE,
  ggtheme  = theme_classic()) +
  xlab("Time (days)") +
  list(ggplot2::scale_color_manual(
          labels = c("None", "Med Imp", "Med Des", "Both"),
          values = c("mediumblue", "goldenrod4", "darkviolet", "midnightblue")),
       ggplot2::scale_fill_manual(
          labels = c("None", "Med Imp", "Med Des", "Both"),
          values = c("mediumblue", "goldenrod", "lightpink", "turquoise")))

 #Frailty and Clustered Cox Models
 cox.LF <- coxph(Surv(duration, status) ~ med_design + med_imp + inter +
                durationDY + multireb + ps_military + ps_economic +
                ps_territory + frailty(regionAdj),
```

```
                data=PSED.clean) # selected by LASSO
cox.LF

cox.LC <- coxph(Surv(duration, status) ~ med_design + med_imp + inter +
                LNdurationDY + multireb + ps_military + ps_economic +
                ps_territory + cluster(regionAdj),
                data=PSED.clean) # selected by LASSO
cox.LC

#Interval Censored Methods
library(survival)
library(icenReg)
library(ggplot2)
library(dplyr)
library(ggfortify)
library(survminer)
library(frailtypack)

#Read CSV with Interval Censored Duration from SAS
PSED <- read.csv("PSEDCLEAN1.csv")

km     <- with(PSED, Surv(icDuration, duration, type = c('interval2'))
          ~ med_imp * med_design)
km.fit <- survfit(Surv(icDuration, duration, type=c('interval2'))
          ~ med_imp + med_design,
          data = PSED)
summary(km.fit, times = c(1, 30, 60, 90 * (1:10)))
```

```
ggsurvplot(km.fit,
  data     = PSED,
  conf.int = TRUE,
  legend   = c(0.7, 0.99),
  ggtheme  = theme_classic()) +
  ggtitle("Log-Rank Survival Curves") +
  xlab("Time (days)") +
  list(ggplot2::scale_color_manual(
          labels = c("None", "Med Imp", "Med Des", "Both"),
          values = c("mediumblue","goldenrod4","darkviolet","midnightblue")),
       ggplot2::scale_fill_manual(
          labels = c("None", "Med Imp",  "Med Des", "Both"),
          values = c("mediumblue", "goldenrod", "lightpink", "turquoise")))


#IC Cox Regression
PSED.ic.frailty <- transform(PSED,
                   status = ifelse(icDuration != durationNonInt, 1, status))
PSED.ic.frailty <- transform(PSED.ic.frailty,
                   icDuration = ifelse(status == 1, durationNonInt - 1,
                    icDuration))


frailtyPenal(SurvIC(icDuration, durationNonInt, status)~ cluster(location) +
   med_imp + med_design + inter + ps_territory + ps_military + multireb,
   data    = PSED.ic.frailty,
   n.knots = 5,
   kappa   = 0.001,
   maxit   = 1000)
```

```
#LASSO Cox for Detailed Power-Sharing Variables
x <- model.matrix( ~ med_design + med_imp + inter + intDY+ LNdurationDY + unpko +
         multireb + pps_cabinet +  pps_sencabinet + pps_nsencabinet +
         pps_parlquota + mps_milcmd +  mps_armyint + eps_company +
         eps_commission + tps_devolution + tps_autonomy + other_proprep +
         other_parlelect + other_preselect + other_referendum +
         other_constitution + other_unresolved, PSED.clean)
y <- Surv(PSED.clean$duration, PSED.clean$status)

#Fit LASSO COX PH MODEL
fit      <- glmnet(x, y, family = "cox", alpha = 1)
allnames <- names(coef(fit)[, ncol(coef(fit))][order(coef(fit)[,ncol(coef(fit))],
                              decreasing=TRUE)])
allnames <- setdiff(allnames, allnames[grep("Intercept", allnames)])
cols     <- rep("gray", length(allnames))
plot_glmnet(fit,
  label = TRUE,
  s     = cv.fit$lambda.min,
  col   = c("coral", "darkorchid4", "blue3", "slateblue1", "goldenrod3",
           "gray27", "magenta"))

cv.fit  <- cv.glmnet(x, y, family = "cox", alpha = 1)
plot(cv.fit)
coef(cv.fit,s = "lambda.min" )

cox.detPS <- coxph(Surv(duration, status) ~ med_design + med_imp + inter +
               LNdurationDY + multireb + pps_cabinet +
               mps_armyint + eps_company + tps_devolution+
```

```
                tps_autonomy + other_proprep+
                other_preselect,
                data = PSED.clean)
cox.detPS # selected by LASSO
coef(cox.detPS)


Survival Random Forest
library(ranger)
library(ggplot2)
library(randomForestSRC)
library(ggRandomForests)
library(ggthemes)
library(survminer)
set.seed(102)
rfsrc.PSED <- rfsrc(Surv(duration, status) ~ med_design + med_imp + inter +
                 intDY + LNdurationDY + unpko + multireb + ps_political +
                 ps_military + ps_economic + ps_territory + ppsPROM_IMP +
                 mpsPROM_IMP + epsPROM_IMP + tpsPROM_IMP,
                 data       = PSED.clean,
                 nsplit     = 1,
                 na.action  = "na.impute",
                 tree.err   = TRUE,
                 importance = TRUE)
plot(gg_vimp(rfsrc.PSED)) +
  ylab("Variable Importance") +
  theme(panel.background = element_rect(fill = "white", colour = "grey50"),
      panel.grid.major   = element_line(colour = "grey50"),
      panel.grid.major.x = element_blank(),
```

```
      legend.position    = "none") +
  scale_color_brewer(palette = "Set1")


plot(gg_rfsrc(rfsrc_PSED), alpha = 0.2, size = 0.55) +
  labs(y = "Survival Probability", x = "Duration (days)") +
  coord_cartesian(ylim = c(-0.01, 1.01)) +
  theme(panel.background   = element_rect(fill = "gray99", colour = "grey50"),
        panel.grid.major   = element_blank(),
        panel.grid.major.x = element_blank(),
        legend.title       = element_blank(),
        legend.position    = "top")+
  scale_color_manual(labels=c("No Failure Observed", "Failed"),
        values = c("mediumblue", "goldenrod4")) +
  scale_fill_manual( labels = c("No Failure Observed", "Failed"),
        values = c("mediumblue", "goldenrod4"))


PSED.sub.strata<-PSED.clean
for(i in 1:79){
  if (PSED.sub.strata[i,1] == 0 & PSED.sub.strata[i, 2] == 0){
    PSED.sub.strata[i, 3]=0 }
  else if (PSED.sub.strata[i,1] == 0 & PSED.sub.strata[i,2] == 1){
    PSED.sub.strata[i, 3] = 1 }
  else if (PSED.sub.strata[i, 1] == 1 & PSED.sub.strata[i, 2] == 0){
    PSED.sub.strata[i, 3] = 2}
  else if( PSED.sub.strata[i, 1] == 1 & PSED.sub.strata[i, 2] == 1){
    PSED.sub.strata[i, 3] = 3}
  }
```

```
set.seed(102)
rfsrc_PSED <- rfsrc(Surv(duration, status) ~ med_design + med_imp + inter +
           intDY + LNdurationDY + unpko + multireb + ps_political +
           ps_military + ps_economic + ps_territory + ppsPROM_IMP +
           mpsPROM_IMP + epsPROM_IMP + tpsPROM_IMP,
  data       = PSED.sub.strata,
  nsplit     = 2,
  na.action  = "na.impute",
  tree.err   = TRUE,
  importance = TRUE,
  splitrule  = "conserve")

plot(gg_rfsrc(rfsrc_PSED, by = "inter")) +
  theme(legend.position = c(0.2, 0.2)) +
  labs(y = "Survival Probability", x = "Time (days)") +
  coord_cartesian(ylim = c(-0.01, 1.01)) +
  scale_color_manual(labels = c("None", "Med Imp", "Med Des", "Both"),
                     values = c("mediumblue", "goldenrod4",
                     "darkviolet","midnightblue"))+
  scale_fill_manual(labels  = c("None","Med Imp","Med Des","Both"),
                    values  = c("mediumblue", "goldenrod", "lightpink",
                    "turquoise")) +
  theme(legend.position = "top",
        legend.title    = element_blank()) +
  theme_classic()

plot.variable(rfsrc_PSED, varnames[1:12])
```

```
#Detailed Power-Sharing RSF


set.seed(102)
rfsrc_PSED <- rfsrc(Surv(duration, status) ~med_design + med_imp + inter +
                intDY + LNdurationDY+ unpko+ multireb + pps_cabinet +
                pps_sencabinet + pps_nsencabinet + pps_parlquota + mps_milcmd +
                mps_armyint +eps_company + eps_commission + tps_devolution +
                tps_autonomy + other_proprep + other_parlelect + other_preselect +
                other_referendum + other_constitution + other_unresolved,
                data       = PSED.sub.strata,
                nsplit     = 2,
                na.action  = "na.impute",
                tree.err   = TRUE,
                importance = TRUE)
plot(gg_vimp(rfsrc_PSED)) +
  ylab("Variable Importance") +
  theme(panel.background   = element_rect(fill = "white", colour = "grey50"),
        panel.grid.major   = element_line(colour = "grey50"),
        panel.grid.major.x = element_blank(),
        legend.position    = "none") +
        scale_fill_brewer(palette = "Paired")


plot(gg_rfsrc(rfsrc_PSED, by = "inter")) +
  theme(legend.position = c(0.2, 0.2)) +
  labs(y = "Survival Probability", x = "Time (days)") +
  coord_cartesian(ylim = c(-0.01, 1.01))+
  scale_color_manual(labels = c("None", "Med Imp", "Med Des", "Both"),
                     values = c("mediumblue", "goldenrod4", "darkviolet",
```

```
                      "midnightblue"))+
  scale_fill_manual(labels  = c("None", "Med Imp", "Med Des", "Both"),
                    values  = c("mediumblue", "goldenrod", "lightpink", "turquoise"))+
  theme(legend.position = "top",
        legend.title    = element_blank()) +
  theme_classic()

#Survival SVM
library(survivalsvm)
library(survival)
PSED.cleaned <- PSED.clean[ , c(2:4, 6:14,16:17, 37,40)]

source("http://bioconductor.org/biocLite.R")
biocLite("survcomp")
survcomp:: concordance.index(survsvm.ranks,
  surv.time  = as.vector(PSED.cleaned$duration),
  surv.event = as.vector(PSED.cleaned$status),
  method     = "noether")

#C-index function for Relative Risk Ranks
concord<-function(rank, duration, status,S){
  count   = 0
  permiss = 0
  concord = 0
  for(i in 1:S){
    for(j in i:S){
     if(i != j & PSED.cleaned[i,14] != PSED.cleaned[j,14] ){
       #Omit when both are end of study right censored at 1826 or 1827
```

```
if(PSED.cleaned[i,15] != 0 & PSED.cleaned[j,15] != 0 ){
   #shorter survival time predicted with lower rank
   if(rank[i, ] >= rank[j, ] & PSED.cleaned[i,14] >= PSED.cleaned[j,14]){
      count   = count + 1
      permiss = permiss + 1}
   else if(rank[i,] <= rank[j, ] & PSED.cleaned[i, 14] <= PSED.cleaned[j,14]){
      count   = count + 1
      permiss = permiss + 1
      }
 #tied rank with higher predicted survival time
   else if(rank[i, ] == rank[j, ] & PSED.cleaned[i, 14] >= PSED.cleaned[j, 14]){
      count   = count + 0.5
      permiss = permiss + 1
      }
   else if(rank[i,] == rank[j, ] & PSED.cleaned[i, 14] <= PSED.cleaned[j, 14]){
      count = count + 0.5
      permiss = permiss + 1
      }
 #tied rank with tied time
   else if(rank[i, ] == rank[j, ] & PSED.cleaned[i, 14] == PSED.cleaned[j, 14]){
      count   = count + 1
      permiss = permiss + 1
    }
   else if(rank[i, ] <= rank[j, ] & PSED.cleaned[i,14] >= PSED.cleaned[j,14]){
      count   = count + 0.5
      permiss = permiss + 1
    }
 }
```

```
      }
    }
  }
 concord = count / permiss
 print(concord)
}

concordSVM <- function(B, d, n, survObj, svmType, kernelType){
  cBootData = data.frame()
  cBootData
  for(i in 1:B){
      d_B <- as.numeric(sample(rownames(d), n, replace = TRUE))

      survsvm.reg <- survivalsvm(survObj,
                 data      = d[d_B, ],
                 type      = svmType,
                 gamma.mu  = c(0.1, 1),
                 opt.meth  = "quadprog",
                 kernel    = kernelType,
                 diff.meth = diffM)
      pred.survsvm.reg <- predict(object = survsvm.reg, newdata = d[-d_B, ])
      survsvm.ranks <- t(pred.survsvm.reg$predicted)
      dOOB <- d[-d_B, ]
      CI <- concord(rank     = survsvm.ranks,
                    duration = dOOB$duration,
                    status   = dOOB$status,
                    S        = length(dOOB))
    dtemp <- as.data.frame(CI)
```

```
    cBootData <- rbind(cBootData, dtemp)
  }
cBootData
}

sObjSVM <- Surv(duration, status) ~ med_design + med_imp + inter + intDY+
                  LNdurationDY + unpko + multireb + ps_political + ps_military +
                  ps_economic + ps_territory +ppsPROM_IMP + mpsPROM_IMP
set.seed(343)
csvm <- concordSVM(B          = 10,
                   d          = PSED.cleaned,
                   n          = 79,
                   survObj    = sObjSVM,
                   svmType    = "regression",
                   kernelType = "add_kernel",
                   diffM      = "makediff1")
csvm <- concordSVM(B          = 10,
                   d          = PSED.cleaned,
                   n          = 79,
                   survObj    = sObjSVM,
                   svmType    = "hybrid",
                   kernelType = "lin_kernel",
                   diffM      = "makediff1")
csvmE <- 1- csvm

#C-index Comparison Across Survival Methods
#C-Index Function For Bootstrap Samples of PSED
library(boot)
```

```
concordE <- function(B, d, n, survObj, coxA, rsf){
  cBootData = data.frame()
  cBootData
  for(i in 1:B){
      d_B        <- as.numeric(sample(rownames(d), n, replace = TRUE))
      cox1       <- coxph(survObj, data = d[d_B, ])
      rfsrc_PSED <- rfsrc(survObj,
                    data       = d[d_B,],
                    nsplit     = 2,
                    na.action  = "na.impute",
                    tree.err   = TRUE,
                    importance = TRUE,
                    ntree      = 1000)
      ApparrentCindex  <- pec:: cindex(list(
                    "Cox 1"  = cox1,
                    "RSF"    = rfsrc_PSED),
                     formula = survObj,
                     data    = d[-d_B, ]
 )
    dtemp <- as.data.frame(ApparrentCindex$AppCindex)
    cBootData <- rbind(cBootData, dtemp)
  }
cBootData
}
sObj <- Surv(duration, status) ~ med_design + med_imp + inter + intDY +
                        LNdurationDY + unpko + multireb + ps_political +
                        ps_military + ps_economic + ps_territory +
                        ppsPROM_IMP + mpsPROM_IMP
```

```
sObj2<-Surv(duration, status) ~ med_design + med_imp + inter +
                LNdurationDY + multireb + ps_military + ps_economic +
                ps_territory
set.seed(7)
c_ind<-concordE(B=100, d=PSED_strata,n=79, survObj=sObj)
c_ind2<-concordE(B=100, d=PSED_strata,n=79, survObj=sObj2)

c_ind$Cox.1 <- 1 - c_ind$Cox.1
c_ind$RSF <- 1 - c_ind$RSF
c_indf <- cbind(c_ind, c_ind2$Cox.1, csvmE)
c_indf$`c_ind2$Cox.1` <- 1 - c_indf$`c_ind2$Cox.1`

par(cex.main = 1.25, cex.axis = 1.25)
par(mar = c(2.5, 4.5, 2.5, 2.5))
boxplot(c_indf,
        ylim  = c(0, 0.60),
        names = c("Cox Baseline", "RSF", "LASSO Cox",  "SVM-Regression-AddK"),
        ylab  ="Concordance Error",
        width = c(1, 1, 1, 1))
```

### B.3 Classification Analysis Code

SAS Code for LDA, QDA, kNN, and Logistic Regression:

```
proc discrim data=psedThesis pool=test;
class durationB;
var med_design med_imp inter intDY LNdurationDY unpko multireb ps_political
ps_military ps_economic ps_territory ppsProm_IMP
```

```
mpsProm_IMP ;
run;
proc discrim data=psedThesis pool=yes crossvalidate;
class durationB;
var med_design med_imp inter intDY durationDY unpko multireb ps_political
ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP;
priors proportional;
run;
proc discrim data=psedThesis pool=no crossvalidate;
class durationB;
var med_design med_imp inter intDY LNdurationDY unpko multireb ps_political
ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP ;
priors proportional;
run;

proc stepdisc data=psedThesis method=backward slstay=0.10 include=0;
class durationB;
var med_design med_imp inter intDY LNdurationDY unpko multireb ps_political
ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP ;
run;
/*STEPWISE LDA*/
proc discrim data=psedThesis pool=yes crossvalidate;
class durationB;
var med_design med_imp inter intDY multireb  ps_military ps_territory;
priors proportional;
```

```
run;

proc discrim data=psedThesis pool=yes crossvalidate;
class durationB;
var durationDY multireb ps_military ps_territory;
priors proportional;
run;
proc discrim data=psedThesis pool=yes crossvalidate;
class durationB;
var intDY durationDY multireb ps_military ps_territory mpsPROM_IMP ps_economic;
priors proportional;
run;

proc discrim data=psedThesis pool=no crossvalidate;
class durationB;
var med_design med_imp inter intDY multireb  ps_military ps_territory;
priors proportional;
run;
proc discrim data=psedThesis pool=no crossvalidate;
class durationB;
var durationDY multireb ps_military ps_territory;
priors proportional;
run;
proc discrim data=psedThesis pool=no crossvalidate;
class durationB;
var intDY durationDY multireb ps_military ps_territory mpsPROM_IMP ps_economic;
priors proportional;
run;
```

```
/*kNN Analysis*/
proc discrim data=psedThesis method=npar k=5 crossvalidate;
class durationB;
var med_design med_imp inter intDY durationDY unpko multireb ps_political
ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP;
priors proportional;
run;

proc discrim data=psedThesis method=npar k=5 crossvalidate;
class durationB;
var med_design med_imp inter intDY multireb ps_military ps_territory;
priors proportional;
run;

proc discrim data=psedThesis method=npar k=5 crossvalidate;
class durationB;
var durationDY multireb ps_military ps_territory;
priors proportional;
run;

proc discrim data=psedThesis method=npar k=3 crossvalidate;
class durationB;
var intDY LNdurationDY multireb ps_military ps_territory mpsPROM_IMP ps_economic;
priors proportional;
run;
proc discrim data=psedThesis method=npar k=5 crossvalidate;
class durationB;
```

```
var intDY LNdurationDY multireb ps_military ps_territory mpsPROM_IMP ps_economic;
priors proportional;
run;

proc logistic data=psedThesis plots=roc;
model durationB =  med_design med_imp inter intDY LNdurationDY unpko multireb
ps_political ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP tpsProm_IMP/ selection=stepwise slstay=0.10 ctable include=3;
run;
proc logistic data=psedThesis plots=roc;
model durationB =  med_design med_imp inter intDY LNdurationDY unpko multireb
ps_political ps_military ps_economic ps_territory ppsProm_IMP
mpsProm_IMP tpsProm_IMP/ selection=backward slstay=0.10 ctable;
run;
```

R Code for Decision Tree, Generalized Linear Mixed Model,
RF, Boosted Trees, and SVM

Decision Tree

```
library(rpart)
library(rpart.plot)
set.seed(32)
PSED.sub <- PSED[ , c(2:16,18:48, 51)]
PSED.rpart <- rpart(durationB ~ . ,
               method = "class",
               control = rpart.control(cp = 0.0, minsplit = 2),
               data    = PSED.sub)
```

```
plotcp(PSED.rpart)

PSED.rpart052 <- rpart(durationB ~ . ,
               method  = "class",
               control = rpart.control(cp = 0.052, minsplit = 2),
               data    = PSED.sub)
PSED.rpart052$variable.importance
plot(PSED.rpart052$variable.importance)
rpart.plot(PSED.rpart052,
  main        = "Binary Duration Classification Tree",
  sub         = "cp = 0.052",
  box.palette = "-Blues")

xvs = rep(c(1:10), length = nrow(PSED))
xvs = sample(xvs)
PSED.rpart052.xval = rep(0, length(nrow(PSED)))
for(i in 1:10){
  train = PSED[xvs != i, ]
  test = PSED[xvs == i, ]
  rp = rpart(durationB ~ .,
           method  = "class",
           data    = PSED.sub,
           control = rpart.control(cp = 0.052))
  PSED.rpart052.xval[xvs == i] = predict(rp, test,type = "prob")[ , 2]}
table(PSED$durationB, round(PSED.rpart052.xval))

GLMM
library(lme4)
```

```
logRegModel1 <- glmer(durationB ~ med_design + med_imp + inter +intDY +
                durationDY+ unpko + multireb + ps_political + ps_military +
                ps_economic + ps_territory + ppsPROM_IMP + mpsPROM_IMP +
                tpsPROM_IMP + (1|regionAdj),
                family = binomial(link = "logit"),
                data   = PSED.clean)
logRegModel1
p <- as.numeric(predict(logRegModel1, type = "response") > 0.50)
table(PSED.clean$durationB, p)
summary(logRegModel1)


set.seed(26)
xvs = rep(c(1:10), length = nrow(PSED.clean))
xvs = sample(xvs)
p   = rep(0, length(nrow(PSED.clean)))
for(i in 1:10){
  train = PSED.clean[xvs != i, ]
  test  = PSED.clean[xvs == i, ]
  logRegModel1 <- glmer(data = train, durationB ~ med_design + med_imp + inter +
                        intDY +durationDY+ unpko+ multireb+ ps_political+
                        ps_military + ps_economic + ps_territory + ppsPROM_IMP +
                        mpsPROM_IMP + tpsPROM_IMP + (1|regionAdj),
                       family = binomial(link = "logit") )
  test
  p[xvs == i] <- as.numeric(predict(logRegModel1,
                type    = "response",
                newdata = test) > 0.50)
}
```

```
table(PSED.clean$durationB, p)


logRegModel2 <- glmer(data = PSED.clean, durationB ~ LNdurationDY + multireb +
                 ps_military +  ps_territory + (1|regionAdj),
                 family = binomial(link = "logit") )
logRegModel2
p <- as.numeric(predict(logRegModel, type = "response") > 0.50)
table(PSED.clean$durationB, p)
summary(logRegModel2)

set.seed(26)
xvs = rep(c(1:10), length = nrow(PSED.clean))
xvs = sample(xvs)

p = rep(0, length(nrow(PSED.clean)))
for(i in 1:10){
  train = PSED.clean[xvs != i, ]
  test = PSED.clean[xvs == i, ]
  logRegModel2 <- glmer(data = train, durationB ~ LNdurationDY + multireb +
                    ps_military + ps_territory + (1|regionAdj),
                    family = binomial(link = "logit") )
  p[xvs == i] <- as.numeric(predict(logRegModel2,
                type    = "response",
                newdata = test) > 0.50)
}
table(PSED.clean$durationB, p)
```

```
logRegModel3 <- glmer(data = PSED.clean, durationB ~ med_design + med_imp +
                      inter + LNdurationDY + multireb + ps_military +
                      ps_territory + (1|regionAdj),
                      family = binomial(link = "logit") )
logRegModel3
p <- as.numeric(predict(logRegModel, type = "response") > 0.50)
table(PSED.clean$durationB, p)
summary(logRegModel3)

set.seed(26)
xvs = rep(c(1:10), length = nrow(PSED.clean))
xvs = sample(xvs)

p = rep(0, length(nrow(PSED.clean)))
for(i in 1:10){
  train = PSED.clean[xvs != i, ]
  test  = PSED.clean[xvs == i, ]
  logRegModel3<-glmer(data = train, durationB ~ med_design + med_imp +
                      inter + LNdurationDY +  multireb +  ps_military +
                      ps_territory + (1|regionAdj),
                      family = binomial(link = "logit") )
  p[xvs == i] <- as.numeric(predict(logRegModel3,
            type    = "response",
            newdata = test)>0.50)
}
table(PSED.clean$durationB,p)
```

RF Models

```
set.seed(98134)
psed.rf <- randomForest(as.factor(durationB) ~ med_imp + med_design +
                          inter + intDY + durationDY + unpko + multireb +
                          ps_political + ps_military + ps_economic +
                          ps_territory + ppsPROM_IMP + mpsPROM_IMP +
                          tpsPROM_IMP, data = PSED)
psed.rf$confusion
psed.rf.xval.class = rep(0, length = nrow(PSED))
psed.rf.xval.prob = rep(0, length = nrow(PSED))
xvs = rep(1:10, length = nrow(PSED))
xvs = sample(xvs)
for(i in 1:10){
  train = PSED[xvs != i, ]
  test  =PSED[ xvs == i, ]
  tempRF = randomForest(as.factor(durationB)~med_imp + med_design + inter +
                        intDY + durationDY + unpko + multireb + ps_political +
                        ps_military + ps_economic + ps_territory +
                        ppsPROM_IMP + mpsPROM_IMP +
                        tpsPROM_IMP, data = train)
  psed.rf.xval.class[xvs == i] = predict(tempRF, test,type = "response")
  psed.rf.xval.prob[xvs == i] = predict(tempRF, test,type = "prob")[ , 2]
  }
table(PSED$durationB, psed.rf.xval.class)
```

Tuned RF Model Mtry = 9 and Ntree = 500

```
set.seed(534)
psed.rf <- randomForest(as.factor(durationB) ~ med_imp + med_design + inter +
                          durationDY + multireb + ps_political +
                          ps_military + ps_economic + ps_territory,
                          data = PSED, ntree = 500, mtry = 9)
psed.rf$confusion
psed.rf.xval.class = rep(0, length = nrow(PSED))
psed.rf.xval.prob = rep(0, length = nrow(PSED))

set.seed(534)
xvs = rep(1:10, length = nrow(PSED))
xvs = sample(xvs)
for(i in 1:10){
  train = PSED[xvs != i ,]
  test = PSED[xvs == i, ]
  glub = randomForest(as.factor(durationB) ~ med_imp + med_design + inter +
                        durationDY + multireb + ps_political+
                        ps_military + ps_economic + ps_territory,
                        data = train, ntree = 500, mtry = 9)
  psed.rf.xval.class[xvs == i] = predict(glub, test, type = "response")
  psed.rf.xval.prob[xvs == i] = predict(glub, test, type = "prob")[ , 2]
}
table(PSED$durationB, psed.rf.xval.class)

GBM: Orginal and Tuned Models

library(gbm)
```

```
library(caret)


Generic GBM
psed.gbm = gbm(durationB ~ med_imp + med_design + inter + intDY + durationDY +
               unpko + multireb + ps_political + ps_military + ps_economic +
               ps_territory + ppsPROM_IMP + mpsPROM_IMP + tpsPROM_IMP,
               distribution = "bernoulli",
               n.trees      = 5000,
               data         = PSED)
table(PSED$durationB, predict(psed.gbm,type = "response", n.trees = 5000))
psed.gbm


set.seed(539)
psed.gbm.xvalpr = rep(0, nrow(PSED))
xvs = rep(1:10, length = nrow(PSED))
xvs = sample(xvs)
for(i in 1:10){
  train = PSED[xvs != i, ]
  test = PSED[xvs == i, ]
  glub = gbm(durationB ~ med_imp + med_design + inter + intDY +
             durationDY + unpko + multireb +ps_political +
             ps_military + ps_economic + ps_territory +
             ppsPROM_IMP + mpsPROM_IMP + tpsPROM_IMP,
             distribution = "bernoulli",
             data         = train,
              n.trees     = 5000)
 # psed.gbm.xvalpr[xvs == i] = predict(glub, newdata = test, type = "prob")[ , 2]
  psed.gbm.xvalpr[xvs == i] = predict(glub,
```

```
                                newdata = test,
                                type    = "response",
                                n.trees = 5000)
}
psed.gbm.xvalpr
table(PSED$durationB, round(psed.gbm.xvalpr))


Tuned GBM
set.seed(124)
fitControl = trainControl(method = "cv", number = 10 )
## tune gbm ##
gbmGrid = expand.grid(interaction.depth = c(3:9),
                      n.trees           = c(25, 30, 40, 45, 50),
                      shrinkage         = c(0.05, 0.06, 0.07, 0.08),
                      n.minobsinnode    = c(3:7))
gbmFit = train( as.factor(durationB) ~ med_design + med_imp + inter + durationDY+
                  multireb + ps_territory + ps_military ,
                  method    = "gbm",
                  tuneGrid  = gbmGrid,
                  trControl = fitControl,
                  data      = PSED)
gbmFit
#n.trees = 50 interaction.depth = 12, shrinkage = 0.05, n.minobsinnode = 10
psed.gbmT = gbm(durationB ~ med_design + med_imp + inter + durationDY +
             multireb + ps_territory + ps_military,
             distribution      = "bernoulli",
             interaction.depth = 5,
             n.trees           = 30,
```

```
            shrinkage         = 0.06,
            n.minobsinnode    = 3,
            data = PSED)
table(PSED$durationB, predict(psed.gbm, type = "response", n.trees = 40))
psed.gbmT

set.seed(124)
psed.gbm.xvalpr = rep(0, nrow(PSED))
xvs = rep(1:10, length = nrow(PSED))
xvs = sample(xvs)
for(i in 1:10){
  train = PSED[xvs != i, ]
  test = PSED[xvs == i, ]
  glub = gbm(durationB ~ med_design + med_imp + inter + durationDY +
           multireb + ps_territory + ps_military, distribution = "bernoulli",
           data              = train,
           interaction.depth = 5,
           n.trees           = 30,
           shrinkage         = 0.06,
           n.minobsinnode    = 3)
  # psed.gbm.xvalpr[xvs == i] = predict(glub, newdata = test, type = "prob")[ , 2]
  psed.gbm.xvalpr[xvs == i] = predict(glub,
     newdata = test,
     type    = "response",
     n.trees = 30)
}
psed.gbm.xvalpr
table(PSED$durationB, round(psed.gbm.xvalpr))
```

```
SVM with Reduction of Variables Selected From Variable Importance in RSF


library(e1071)


tunedParams <- tune.svm(as.factor(durationB) ~  med_design + med_imp + inter +
                    multireb + ps_military + ps_territory + LNdurationDY,
                    data  = PSED,
                    cost  = 10^(-3:3),
                    gamma = 10^(-4:3))
summary(tunedParams)


#Tuned Parameters: Cost=100  Gamma=0.001


set.seed(431)
xvs = rep(1:10, length = nrow(PSED.clean))
xvs = sample(xvs)
for(i in 1:79){
  train = PSED.clean[xvs != i, ]
  test = PSED.clean[xvs == i, ]
  svmtemp = svm(as.factor(durationB) ~ multireb + med_design + med_imp + inter +
                    ps_military + ps_territory + LNdurationDY,
                    data = train,
                    cost = 100,
                    gamma = 0.001)
  PSED.svm.pred[xvs == i] = predict(svmtemp, test)
}
```

```
PSED.svm.pred[]

table(PSED.clean$durationB, PSED.svm.pred)
```