Statistical Learning for Duration Analysis, with application to missing persons data in Montana

Thesis · April 2021

DOI: 10.13140/RG.2.2.14156.63361

Emmanuel Sarfo Fosu
Baylor University

Statistical Learning for Duration Analysis, with application to missing persons data in Montana

Emmanuel Sarfo Fosu

Department of Mathematical Sciences

Montana State University

April 30, 2021

A writing project submitted in partial fulfillment

of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

EMMANUEL SARFO FOSU

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Dr. Mark C. Greenwood
Writing Project Advisor

Date

Dr. Mark C. Greenwood
Writing Projects Coordinator

Abstract

Statistical learning was introduced in the 1960s with its purpose being estimation of theoretical functions, until the 1990s where technological advancement helped improved this area of study to a practical algorithm tool. This area of statistics is thoroughly developed for quantitative or categorical responses. However, it has also been extended to survival data. Most survival data analyses use methods such as the Cox Proportional Hazard model in their analysis, but the proportional hazards (PH) assumption is a strong assumption of this method and is often violated. The statistical learning tool Random Survival Forests (RSF) was suggested to be a good alternative for the Cox proportional hazards model when the PH assumptions are violated and they also can have better predictive performance by incorporating interactions and nonlinear functions of quantitative predictors.

The objective of this study was to apply Random Survival Forests (RSF) to time to resolution of missing persons case in Montana from 2017 to 2019 and compare this statistical learning method to conventional statistical methods used in survival analysis.

The study findings indicate that the Random Survival Forests (RSF) model has a better prediction error than the Cox model and Kaplan Meier curves based on Brier scores but the difference is minimal. Using the C-Index as a prediction evaluation, the Random Survival Forests (RSF) model performed better in the initial time but was overtaken by the Cox model at later time points. From the RSF model, age, region, cases reported on reservation, and race are seen to be important for the time to resolution of a missing person case.

Acknowledgments

I would like to thank my advisor, Dr. Mark Greenwood, for his time dedicated to this project, corrections, and directions. This work would not have been successful without his continuous guidance.

I also thank Tina Chamberlain from the Montana Department of Justice for allowing me to use their data set for my project. A special thanks to Greta Linse, Assistant Director of SCRS, for her immense contribution to getting the data all cleaned up for this project.

My sincere gratitude goes to the Statistics faculty in the Montana State University Department of Mathematical Sciences for guiding me throughout my Master's degree and helping me grow as a Statistician.

Finally, I place on record, my sense of gratitude to all who have directly or indirectly contributed to this project.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background of Statistical learning

Statistical learning theory was introduced in the 1960s where it was primarily used for estimation of theoretical functions from a given set of data. There were new types of learning algorithms such as support vector machines (SVM), support vector regression (SVR), and support vector clustering (SVC) proposed in the 1990s which improved and transformed the statistical learning theory to a practical tool for estimating multidimensional functions (Vapnik, 1999).

Statistical learning theory involves studying, understanding, and accurately modeling a predictive function for complex datasets. Statistical learning can be classified as supervised or unsupervised learning. Supervised learning involves problems of regression and classification based on one or more inputs and a known response variable. With unsupervised statistical learning, the goal is learning about the patterns and structure of the data (Mishra, 2017). Statistical learning plays a key role in many areas of science, technology, engineering, and medicine. In its operations, it employs split sample and/or cross-validation techniques where data are usually split into training and test sets or sometimes split more than once in the case of cross-validation. The training dataset is used for learning the relationship and the accuracy of model is tested using the test dataset.

As time has gone by, statistical learning tools which previously have been developed to model problems with categorical and quantitative responses have been extended to time to event or survival response data. This rapid progress in statistical learning has been attributed to the increasing availability of powerful and user-friendly softwares (James et al., 2013). It is of no doubt that this field in statistics would continue to expand and become an even more essential tool for non-statistical fields.

1.2 History and Background of Survival or Duration Analysis

Survival (duration) analysis is the analysis of time between entry to a study and an event of interest. When survival analysis was introduced in the 17^{th} century, it was only focused on time from treatment until death, hence the name (Camiller, 2019). Survival analysis received a major boost in 1958 and 1972 following the great contributions of Kaplan and Meier (1958) in the estimation of survival probabilities and hazard rates, and comparing the relative forces of mortality of two groups and Cox (1972) for the proposal of the proportional hazards model which explains how risk changes over time based on predictor variables. Other deserving mentions in the history of survival analysis are Vaupel et al. (1979) and Hougaard (1995), for the introduction of shared and unshared frailty models.

These statistical methods do not follow the standard linear model as they incorporate the information associated with subjects who did not experience the event either by end of study or were lost to follow-up as subjects withdrew from the study before experiencing the event of interest, with both creating censored information. Therefore, it is very important to define the event of interest and the timeline of the study in survival/duration analysis. That is, the beginning and endpoint of the study must be stated so that subjects of the study are on equal grounds from the start for the process to be appropriately analyzed (Kartsonaki, 2016). The quest for accuracy is very fundamental in all aspects of life, especially in health sciences. In most clinical settings, the time between exposure and event is very useful. Survival/duration analysis has been an important inferential statistical tool used in clinical studies. Over the years, survival/duration methods have been extended to non-clinical settings like business sales, insurance, evaluating product reliability, engineering, economic activities, etc. Survival/-duration analysis is able to assess evidence for risk factors, estimate their impacts, and rank them by their importance during prediction of survival times (Camiller, 2019).

The objectives of survival/duration analysis include the analysis of patterns of event times, the comparison of distributions of survival times in different groups of individuals, and examining whether and by how much some factors affect the risk of an event of interest. In survival/duration analysis many different regression modeling strategies can be applied to predict the risk of future events, and three different approaches are discussed here: Kaplan-

Meier curves, Cox Proportional Hazard models, and Random Survival Forests.

1.3 Study Objectives

The conventional statistical method for the analysis of right censored survival data is the Cox proportional hazards (Cox PH) model which is discussed in detail below. This model analyzes the effects of covariates on survival time and has widely been used because of its flexibility for modeling time-to-event with censoring. The proportional hazards (PH) assumption is that the hazard ratio is constant over time. PH, however a strong assumption of this method and is often violated (Nasejje et al., 2017).

The statistical learning tool, Random Survival Forests (RSF), which is an extension of the conventional Random Forests (RF) method for survival data has been suggested to be a good alternative for the Cox proportional hazards model when the PH assumptions are violated and can also have better predictive performance (Ishwaran et al., 2008). Recent studies have shown that RSF may have some flaws with regards to favoring covariates with many split-points.

This paper explores the application of statistical learning for time-to-event data and compares these statistical learning tools compare against the conventional statistical methods used in analyzing time-to-event data, with an application to time to case resolution of missing persons in Montana from 2017 to 2019. There were $n = 5,566$ total unique events reported and these were randomly split into training ($n = 3,897$) and test data sets ($n = 1,669$) for the following analyses.

2 Methodology

2.1 Survival/Duration Analysis

2.1.1 Censoring

Survival/duration analysis is concerned with time to an event of interest. However, there are often some subjects who may not be observed on the event of interest within the expected duration of the study. This is the phenomenon of censoring. Censoring may come about in the following ways:

1. Participant withdraws from the study.
2. Participants do not experience the event of interest before the study ended.
3. Participant is lost to follow-up during the study period, so the researcher does not know if the event has occurred.

There are different types of censoring in survival/duration analysis. Three of them are explained below:

Right Censoring: This is the most observed and easiest type of censoring in the analysis of time-to-event data. Right censoring occurs when a participant starts at the beginning of the study (time=0) and exits before the event of interest occurs. That is, either there is a withdrawal from the study before the event of interest occurs or the event of interest was not observed before the study ended. The last time of known status or the end time of the study is the last information available for the subject.

Left Censoring: Left censoring occurs when we cannot observe the time when the event occurred, or the event of interest occurred at an unknown time before the start of the study.

Interval Censoring: In interval censoring, the time to the event of interest is only known to lie within a time period instead of being observed exactly. Kartsonaki (2016) also explains interval censoring as where the participant's time to event is between two time points but the exact point not known.

For this study, we will only consider right censoring for the application of the methods. That is, censoring may occur when a subject of a reported case has not been found or resolved before the study ended. Censoring is also assumed to be non-informative so the censoring time of each subject is independent of their failure time. This enables the analysis to be devoid of any bias.

2.1.2 Survival Function

Let $T \geq 0$ be a continuous random variable representing the survival time with a probability density function, $f(t)$, and a cumulative distribution function, $F(t) = Pr(T \leq t)$. Then with some observed time t, the survival function $S(t)$ is the probability that a participant survives beyond time t. It is given as

$$S(t) = 1 - F(t) = P(T > t). \tag{1}$$

Note that throughout the following discussion survival time is used interchangeably with duration or time to event.

2.1.3 Hazard Function

The hazard function is the instantaneous rate at which an event of interest occurs for participants which have active cases at time t. It can also be defined as the ratio of the probability density function and the survival function. The hazard function is given as:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}, \tag{2}$$

where δt is width of the interval.

Based on Equation (2), the hazard function can be transformed using the definition of conditional probabilities and derivatives,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \cdot \frac{1}{S(t)}.$$

Therefore the hazard function can also be defined as:

$$h(t) = \frac{f(t)}{S(t)}. \tag{3}$$

2.1.4 Cumulative Hazard Function(CHF)

Applying derivative laws to Equation (1), we can write Equation (3) as:

$$h(t) = -\frac{d}{dt} \log S(t). \tag{4}$$

Then after taking the integral on both side of Equation (4),

$$S(t) = e^{-\int_0^t h(u)du}. \tag{5}$$

The integral function in Equation (5) is called the *Cumulative Hazard Function (CHF).* It is formally denoted as :

$$H(t) = \int_0^t h(u)du. \tag{6}$$

The CHF is the total amount of the risk encountered from time 0 to t. From Equation (5), the relationship between the CHF and survival function is observed. The equation suggests that when there is higher hazard, the survival probability becomes lower.

2.2 Conventional Survival Analysis

2.2.1 Kaplan-Meier Estimation

In survival/duration analysis, one of the purposes is to estimate the hazard and survival functions, which are used to summarize the survival data. Kaplan-Meier estimation is one of the methods for estimating the survival function. Usually survival times are skewed to the right and so it becomes difficult to find parametric distributions that match the observed distributions. A non-parametric approached is useful for estimating the survival function, such as was proposed in Kaplan and Meier (1958).

One unique feature about Survival data is censoring. Since we need to incorporate censored information into the estimation, the Kaplan-Meier method is able to account for the censoring of the data and even make use of the information from these subjects up to the time when they are censored. The Kaplan-Meier estimator is defined as the probability of surviving in a given length of time while considering time in many small intervals (Altman, 1992). There are three assumptions as in described Koletsi and Pandis (2017), used in this analysis. They are:

1. Censored subjects have the same likelihood of surviving as those who continue to be studied at any time.

2. The survival probabilities are the same for subjects recruited early and late in the study.

3. The event happens at the time specified.

Not listed as an assumption in Koletsi and Pandis (2017) but also implicit in these methods is that observations are independent of one another. The Kaplan-Meier survival probability at any particular time t is calculated by the formula:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i}\right), \tag{7}$$

where $\hat{S}(t)$ is the probability that a subject survives longer than time t, n_i represents the number of subjects at risk prior to time t, d_i represents the number of the event of interest at time t, and t_i is a time when at least one event happened.

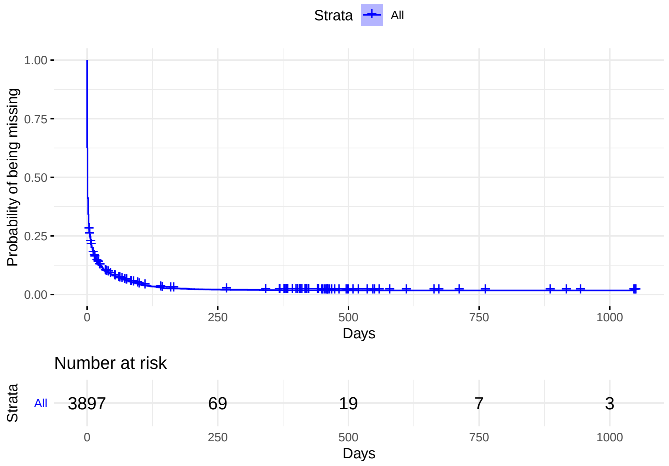


Figure 1: Kaplan-Meier curve for duration of being missing for missing cases reported in Montana, 2017-2019 for the $n = 3,897$ training observations.

The estimated probability $(\hat{S(t)})$ is a step function that changes value only at times of events. It is also possible to compute confidence intervals for the survival probability at each t, although these are not included. Figure 1 demonstrates Kaplan-Meier estimation for the $n = 3,897$ used in the training data estimated using the survival package (Therneau, 2020) in the statistical software R (R Core Team, 2021). The Kaplan-Meier curve suggests that the estimated probability of remaining missing drops quickly in the initial few days after a missing person is reported.

2.2.2 Cox Proportional Hazard model

The Cox Proportional Hazard model (CPH) is a commonly used model in survival analysis. CPH is a semi-parametric model because the component associated with the covariates assumes an underlying parametric distribution while the baseline hazard function makes no parametric assumption regarding the nature of the hazard function. The event of interest may depend on several covariates, Cox (1972) developed this model to investigate the association between the survival time and one or more covariates or determine which combination of potential predictor variables impact the form of the hazard function, and how they impact it. The Cox model hazard function, $h(t)$, can be estimated as:

$$h(t) = \lambda_0(t)e^{\{\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p\}}, \tag{8}$$

where $\lambda_0(t)$ represents the baseline hazard function, $x_1, x_2, ..., x_p$ are the covariates, and $\beta_1, \beta_2, ..., \beta_p$ are the coefficients from the model that relate the covariates to changes in the hazard (and survival rates).

From Equation (8), the hazard at time t is the product of the baseline hazard function (all covariates values are 0) and the parametric component (exponentiated linear function of covariates). Since the Cox proportional hazards model is a semi-parametric model, it depends on some assumptions to provide valid inferences:

1. The proportional hazard assumption, which states that the proportional hazard must be constant over the time t.

2. Censoring must be non-informative.

3. Observations are independent of one another.

4. Quantitative predictors are modeled correctly (transformation or polynomials are not needed).

5. Interactions not included in the model are not needed.

For the application to missing persons data, all assumptions might be problematic or violated, but moving to statistical learning methods can alleviate concerns about assumptions 1, 4, and 5. Assumption 2 could be violated due to the nature of certain events where censoring might be related to the characteristics of the event. Assumption 3 is violated because the same subject can be (and often is) observed multiple times in the data set with only $3,254$ unique subject responsible for the $n = 5,567$ total reports and subjects missing over long periods of the study can't be reported missing again. Despite these potential limitations, the modeling of missing duration times can provide useful information about the characteristics of the cases.

2.3 Statistical Learning application to Survival Analysis

2.3.1 Random Forests

Random forests are a supervised learning algorithm that build ensembles of decision trees together. One unique feature about this model is that it adds additional randomness to the model, while growing each of the trees in two ways. In each tree, the random forest model searches for the best feature among a random subset of features as the best splitting candidate to split nodes and each tree is built on a bootstrap sample of the data as shown in Figure 2. Random Forests have been known to perform very well because it operates on a fundamental concept which Yiu (2019) states as *"A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models."*

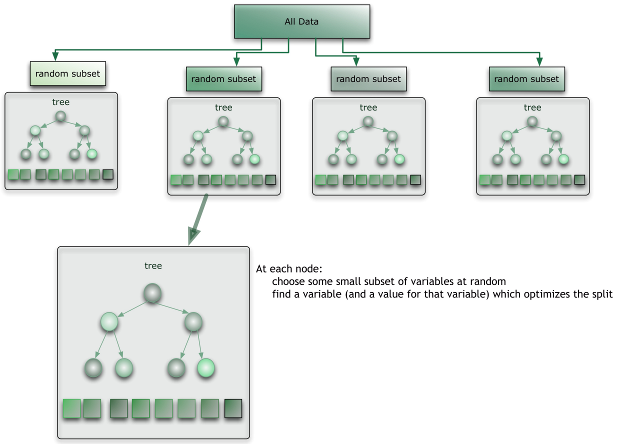

All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 2: A pictorial representation of the mechanism of Random Forests from Benyamin (2012).

The bootstrapping of observation for each tree and subsampling of predictors at each node provides a suite of more independent trees to combine in ensemble forest (Hastie et al., 2001). This ensemble tree model was initially developed to model continuous and discrete response variables, which solves classification and regression problems. However, Breiman (2001) stated that this mechanism can be extended to censored data as well.

2.3.2 Random Survival Forest

The idea of extending the random forests to right-censored data was developed by Ishwaran and Kogalur (2007). Their Random Survival Forest (RSF) is an ensemble tree method for analyzing right-censored survival data (Ishwaran et al., 2008). This statistical approach to modeling survival data is a non-parametric model and therefore relaxes some of the assumptions set by the Cox PH model. Random Survival Forest (RSF) is able to handle high dimensional covariate time-to-event data and model time-to-event data with nonlinear relationships for predictor variables, and also identify types of interactions (Bou-Hamad et al., 2011) without manually specifying them. The analysis of the RSF in this study is done using the randomForestSRC R package (Ishwaran and Kogalur, 2020).

Algorithm Process

The algorithm of the Random Survival Forest is described by Ishwaran et al. (2008) as:

- Draw B bootstrap samples from the original data set. Each bootstrap sample leaves out the out-of-bag (OOB) data (approximately $\frac{1}{3}$).
- For each B sample, a survival tree is grown where at each node we randomly select $m \leq p$ predictors as candidate splitting variables, ($m =$ number of candidate splitting variables, $p =$ total predictor variables). The particular variable in m used to split the node should be the one that maximizes the survival difference between daughter nodes.
- Grow each tree to full size until terminal nodes have no less than d_0 unique events ($d_0 =$ nodesize).
- The cumulative hazard function (CHF) for each tree is calculated and it is averaged over B survival trees to obtain the ensemble CHF.
- The prediction error for the ensemble CHF and variable importance are calculated using OOB data.

For a terminal node of a tree, which is represented here with k, the cumulative hazard function (CHF) estimate is estimated using the Aalen (1978) estimator:

$$\Lambda_k(t) = \sum_{t_{l,k \leq t}} \frac{d_{l,k}}{W_{l,k}}, \tag{9}$$

where $d_{l,h}$ is the number of deaths at time $t_{l,h}$ and $W_{l,h}$ is the individuals at risk at time $t_{l,h}$.

Once we have the CHF for one tree, we then find the ensemble CHF, which is based on averaging over the B bootstraps. We compute the ensemble CHF for bootstrap and out-of-bag (OOB) error in order to help us assess our model's predictive ability and also to assess variable importance in the model. The ensemble CHF for out-of-bag (OOB) data is

$$\Lambda^O(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b}\Lambda_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}},$$

where $I_{i,b} = 1$ if i is an OOB sample and $I_{i,b} = 0$ otherwise, and $\Lambda_b(t|x_i)$ is the CHF of a tree from the b^{th} bootstrap sample. The ensemble CHF for the bootstrap sample is

$$\Lambda^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} \Lambda_b(t|x_i).$$

Using the idea of Equation (5), the ensemble survival function for the RSF is calculated as:

$$S(\hat{t}|x_i) = e^{-\Lambda^*(t|x_i)}.$$

Splitting rule

One important component of building ensemble trees is the split rule. For computations in the `randomSurvivalForest` by Ishwaran and Kogalur (2020), there are four splitting rules for the random survival forest. For this paper, only 2 of these splitting rules are considered. They are the log-rank splitting rule and log-rank score splitting rule. As described by Segal (1988), the `log-rank` splitting rule works on the idea that the best split at a node k is using the candidate splitting variable that gives the largest log-rank statistic between the two daughter nodes. The limitation of this split rule is that it is biased towards variables with many split points.

Hothorn and Lausen (2003) explains the `log-rank score` splitting rule as a method which uses the standardized log-rank statistic to split the nodes. That is, the candidate splitting variables are ordered, the ranks of each survival time are computed, and then it uses the log rank score to make the split decision. This can provide better performance for balancing variables with many split points that are possible.

Variable Importance (VIMP)

Variable importance measures (VIMP) provide methods for understanding the relative importance of predictors in a model. In particular, they show the predictive performance of each variable in the model. Breiman (2001) and Ishwaran and Kogalur (2007) calculate the variable importance (VIMP) as the difference between the prediction error for the new randomized predictor ensemble and the prediction error for the original ensemble. A large VIMP value for a predictor shows that the variable is important and has high predictive power, whereas for zero or negative VIMP values, the variables are considered less important. It should be noted that a VIMP of 0 implies a variable has no effect on the prediction error. That is, growing the forest with or without a variable would not change the prediction error.

2.3.3 Prediction error in survival models

For predictive models, there is the need to evaluate the method to ensure that prediction is done accurately. In this paper, we consider two ways of evaluating the survival model. They are the Concordance index (C-index) and prediction error using Brier scores.

Concordance Index (C-Index)

The Concordance Index (C-Index) estimates the probability of correctly predicting the survival time of a pair of observations such that the subject with the longer survival time has the higher probability of survival predicted by the model. The C-Index depends on time for its evaluation and takes into account censoring. The following outline was given by Ishwaran et al. (2008) to calculate the concordance index:

1. Using the entire data set, form all possible pairs of cases.

2. Drop all pairs with censoring in the observation with the shorter survival time and pairs

with the same survival time unless at least one has experienced the event of interest. The remaining pairs are the total number of permissible pairs.

3. Each pair of cases are scored based on the following:
   - A pair receives a score of **1** if they have unequal survival time and the shorter survival time is predicted to have a worse outcome or **0.5** otherwise.
   - A pair receives a score of **1** if they have equal survival time and both have not experienced the event, the predicted outcome is worse for the observation with the observed event or **0.5** otherwise.
4. Concordance is the sum of these scores over the count of all permissible pairs. The C-index, C, is defined as C = Concordance/Permissible.

The C-Index is plotted against time to show the prediction performance of the model with time. A C-Index closer to 1 suggest a better prediction and a C-Index = 0.5 is equivalent to randomly guessing.

Brier scores

Survival data often include censoring and any model assessment metrics need to account for this incomplete information on some records. Another method for evaluating survival models is prediction error, which can be calculated using the Brier score. Brier scores are calculated as the squared difference between observed event status and the predicted survival probability at time t for observations in the test set based on the forest trained on a training data set. Mathematically, Nasejje et al. (2017) define the Brier score as:

$$BS(t) = \frac{1}{N}\sum_{k=1}^{N}[0 - \hat{S}(t|X_k)]^2\frac{I(t_k \leq t, \delta_k = 1)}{\hat{R}(t|X_k)} + [1 - \hat{S}(t|X_k)]^2\frac{I(t_k > t)}{\hat{R}(t|X_k)}. \tag{10}$$

where N is the number of subjects in the data set, $\hat{S}(t|X_k)$ is the predicted survival probability using the training data set for subject k at time t, $\hat{R}(t|X_k)$ is the estimate of the conditional survival function of the censoring times, and δ_k is the event status of subject k.

The Brier scores are executed using `pec` function in the pec package in R (Mogensen et al., 2012). This method evaluates prediction errors on the training and test data set. When a Brier score prediction error is 50% or higher, then the prediction is just as randomly guessing, and over 50% suggests it is worse than guessing.

3 Application to Missing Persons data

3.1 Data

The methods discussed previously are applied to data of missing persons from the Montana Department of Justice, reported from 2017 to 2019. The original data set contains $n = 5,567$ reported cases in the state with 20 variables taken on every report. For this study, the total sample to be used is 5,566, divided randomly into train ($n = 3,897$) and test ($n = 1,669$) datasets. The event of interest is the time to be found for each of the cases reported.

- IntC: the number of days until a person is found or missingness is resolved, or end of study period on December 31, 2019.
- Status: Status of censoring ($0 :$ censored, $1 :$ the person is found by December 31, 2019).

We focused on 7 predictors in this study and they are explained below:

- Reserve: Reporting case on reservation (Yes) or not reporting on reservation (No). Cases without a reservation reported were coded as "No"as well.
- Age: The age of the missing person in years at the time of reporting.
- Gender: Whether the missing person is a male (M) or female (F).
- Race: The race of the missing person labeled as Asian, Black, Indigenous, White, or Unknown.
- Circumstance: The circumstances under which the person was reported missing (runaway, fed mandate, noncustodial parent, or unknown). Cases without circumstance reported were coded as "unknown". Circumstance had one observation with a unique level that was omitted from this study.
- Region: Counties are groups into regions based on where they are located in the state. This variable was created using information on state region as defined by Montana Pollution Prevention Program (2021), (NorthCentral, NorthEast, NorthWest, SouthCentral, SouthEast, or SouthWest).

- Year : The year in which case was reported (2017, 2018, or 2019).

3.2 Data Analysis

3.2.1 Kaplan Meier (KM) curves for duration of being missing

The Kaplan-Meier survival curve produces a step function that changes value only at the time when a missing person is found. This plots the estimated probability of being missing against time. The `ggsurvplot` function in the Survminer R package (Kassambara et al., 2020) is used to produce these Kaplan-Meier curves. The probability of being missing is compared for levels of each covariate using a log-rank test statistic, which finds evidence against the null hypothesis of no differences in the duration of being missing.

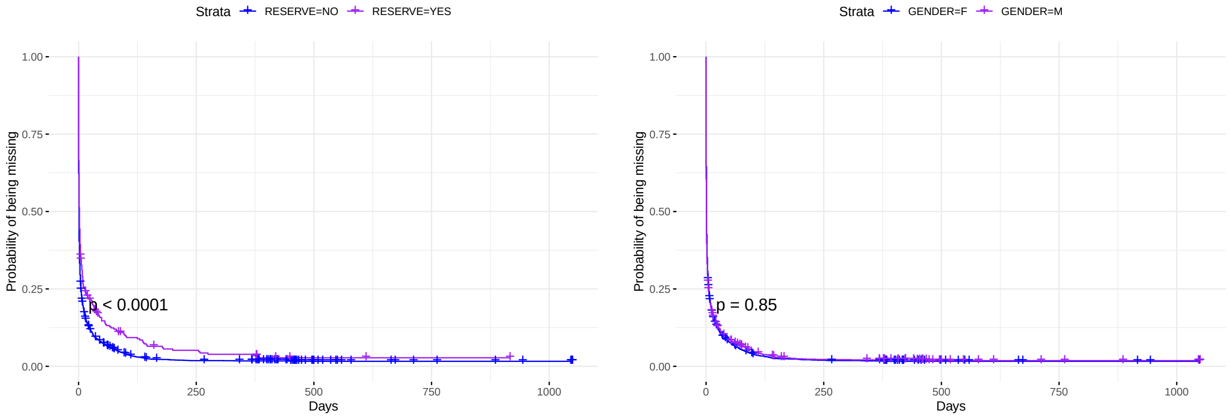


Figure 3: Kaplan-Meier curve for Reservation (left panel) and Gender (right panel).

Figure 3 shows a Kaplan-Meier curve for missing time length for reservation (left) and gender (right). It is observed that cases not reported on reservations have higher probability of being found than cases reported on reservations. The log-rank p-value < 0.0001 suggests that there is a strong evidence against the null hypothesis of no difference in distribution between reporting on/off reservation and this suggests that there is some difference in the duration of being missing between the groups. Also looking at the Kaplan-Meier curve for gender (Figure 3 right panel) right, the log-rank p-value $= 0.85$ suggests that there is a little to no evidence that there is some difference in the duration of being missing between the two groups.

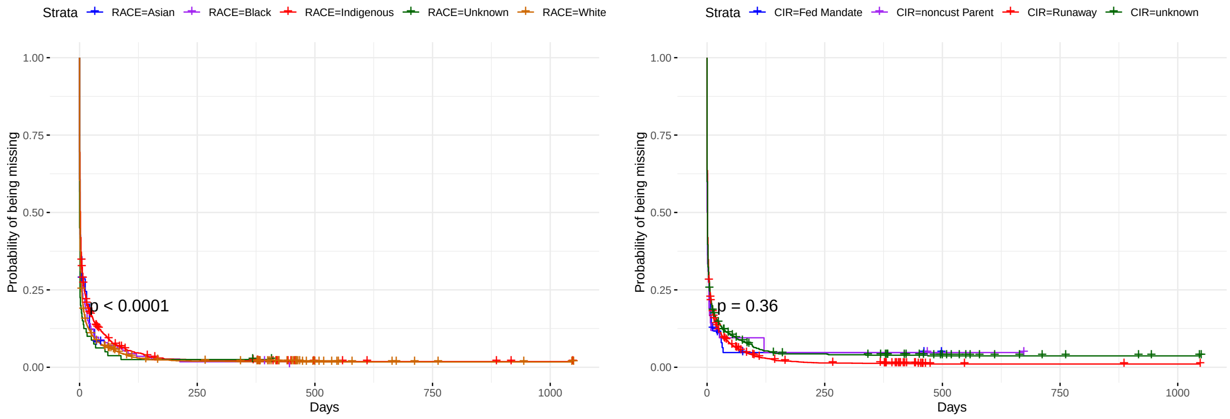


Figure 4: Kaplan-Meier curve for Race (left panel) and Circumstance (right panel).

Figure 4 shows a Kaplan-Meier curve for race (left) and circumstances of report (right). The log-rank p-value < 0.0001 for the race groups suggest that there is a very strong evidence that there is some difference in the duration of being missing. Interestingly, the Kaplan-Meier plot does not suggest large differences among the groups but the small p-value may be attributed to the large sample size providing high power to detect small differences in the duration curves. The log-rank p-value $= 0.36$ for circumstance suggests that there is a weak evidence that there is some difference in the duration of being missing between the circumstances under which a person was reported missing, but the unknown group is a large proportion of the dataset.

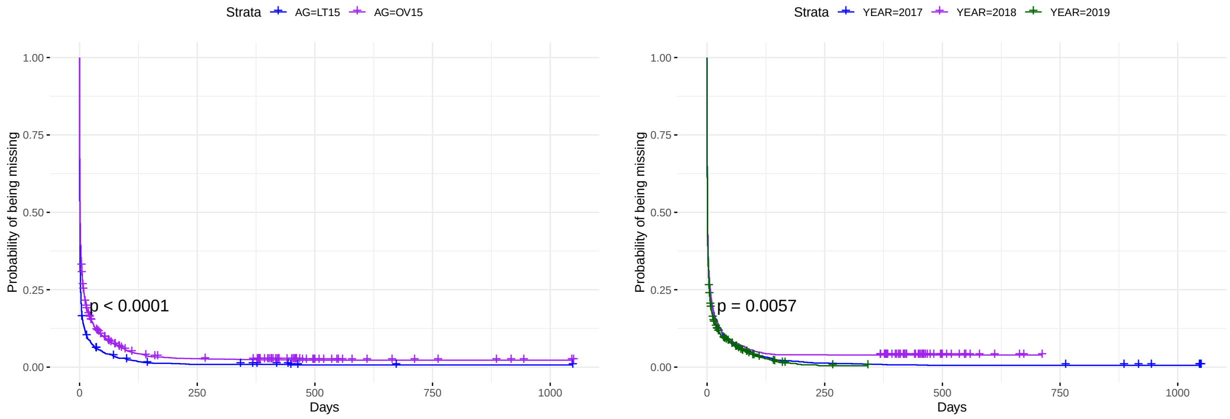


Figure 5: Kaplan-Meier curve for Age (left panel) and Year (right panel).

Looking at Figure 5 it shows a Kaplan-Meier curve for age (left) and year (right). To explore Kaplan-Meier curves for age, it was divided into < 15 (younger) and $\geq$ (older). The age Kaplan-Meier plot shows that subjects over 15 years tend to be missing for longer than the younger ones. The log-rank p-value < 0.0001 suggests that there is very strong evidence that there is difference in the duration of being missing between the over 15 years and under 15 years groups. The year groups also showed that missing duration were longer in 2018 than the other two years. The log-rank p-value $= 0.0057$ suggests that there is strong evidence of a difference in duration of being missing for at least one of the years. It is unclear why this difference might be present and more years should be included in the data set to see if similar results are found or if this is just due to year-to-year variation in results. Comparison to 2020 results during the COVID-19 pandemic will be of a particular interest for a future similar comparison.

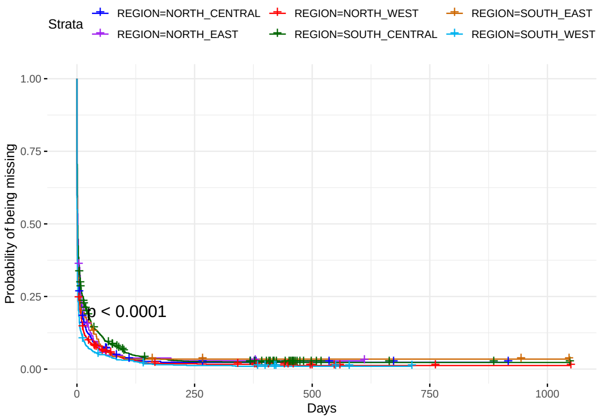


Figure 6: Kaplan-Meier curve for Region.

There is very strong evidence based on the log-rank p-value < 0.0001 that there is some difference in the duration of being missing among the regions based on the results in Figure 6. The differences are not easily seen on the Kaplan-Meier curve shown in Figure 6 so the large size of the data set may be contributing to the small p-value here. Variation across regions of the state could be related to presence/absence of other variables being considered. This is addressed in more complex analyses that follow, where reservation status for the report and other characteristics of the subjects are accounted for and might be related to differences observed by region.

3.2.2 Cox proportional hazard (Cox PH) model for duration of being missing

The Cox PH model is fitted to simultaneously assess the association between the 7 covariates and duration of being missing. It estimates the hazard function which in this study is the probability of being found at time t_i given the subject was missing at and beyond the time t_i. The measure of risk provided for each variable is the hazard ratio (HR).

- When **HR= 1**, there is no effect of the covariate on the rate of being found.
- When **HR<1**, this suggests that the covariate contributes to an increase in duration of being missing.
- When **HR> 1**, this suggests that the covariate contributes to the reduction of a person's duration of being missing.

The ANOVA table in Table 1 shows very strong evidence that the race of the individual, region where the case was reported, age of the individual, and year when case was reported have impacts on the time to resolution of missing case conditional on all other terms in the model. Also, there is strong evidence that the circumstances and reservation have impacts on the time to resolution of missing case. However, there is weak evidence that the gender of the missing person has impact on the time to resolution after adjusting for the other variables. The information from Table 1 indicates that all the variables except for gender are needed in the Cox PH model conditional on each other. However, the gender variable is retained in the model to compare results to our later statistical learning approach. The results show that there are some differences in the hazard ratios in the circumstance, region, age, reservation, and year after controlling for all other variables. The Cox PH model is fitted with the 7 covariates using the `coxph` function in the `survival` package in R (Therneau, 2020).

	LR	Chisq Df	Pr(>Chisq)
GENDER	0.465	1	0.495
RACE	18.841	4	0.001
CIRCUMSTANCE	8.908	3	0.031
REGION	96.045	5	< 0.001
AGE	54.877	1	< 0.001
RESERVE	5.386	1	0.020
YEAR	12.043	2	0.002

Table 1: Analysis of Deviance Table (Type II tests).

The output shown in Figure 7 shows the hazard ratios of the covariates and their respective 95% confidence interval. Each result is conditional on the other predictors in this model. The results in Figure 7 show that the hazard rate for time to resolution for black and white individuals reported missing is 1.01 times higher compared to Asian (HR=0.9). Additionally, Indigenous race subjects are less likely to be found than Asian. It is seen that all circumstances under which a person was reported missing are less likely to be found when compared to the federal mandate group. Also, a reported case in the Southwest region 27% higher hazard rate to be resolved than a case in NorthCentral region. Cases reported in Northeast and SouthCentral region of the state may take a longer time to be resolved when compared against NorthCentral. For a one year increase in the age of the missing person, the hazard ratio is 0.99 times lower suggesting that older missing people are more likely to be missing longer. A case which was reported on a reservation has a 15% lower hazard ratio to be resolved compared to those not reported on a reservation, also suggesting longer missing times for reservation reported cases. Using 2018 as the reference year, it is observed that cases reported in years 2017 and 2019 are likely to be resolved earlier when compared with 2018.

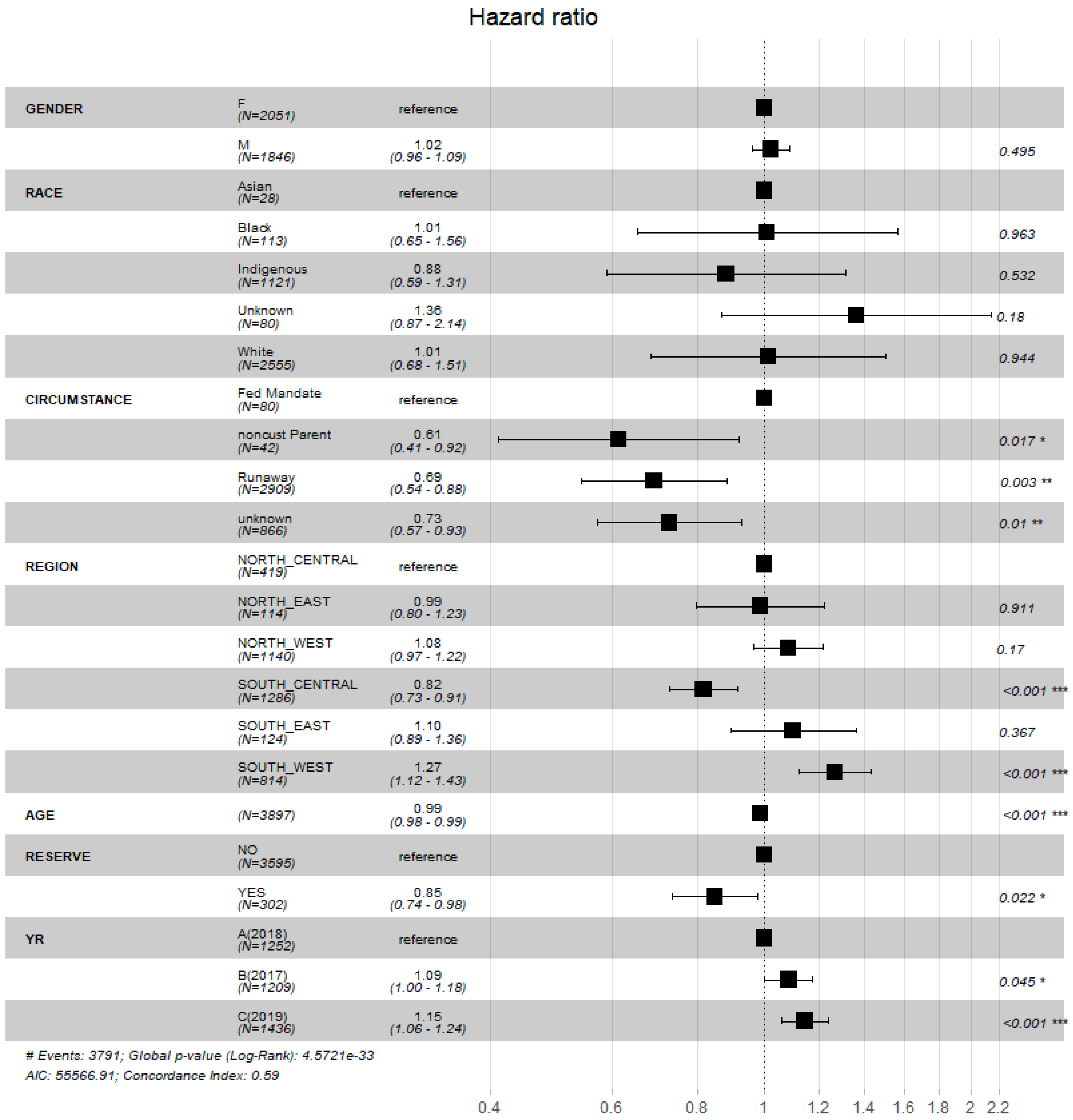


Figure 7: Cox PH output: Reference level of categorical covariates denoted `reference` and have estimated HR of 1.

3.3 Random Survival Forest (RSF)

A random survival forest (RSF) uses independent bootstrap samples to grow an ensemble of survival trees that split each node based on the covariates and a criterion involving time to resolution of a missing case and censoring status information. We predict the overall time to resolution using the 7 covariates. We use the random survival forest model where 1,000 trees are grown with the training data using the logrank and logrankscore splitting criteria, which are then assessed in the test dataset.

The RSF model is also fitted on permuted data to serve as a reference model to compare the real data set to one with no relationship between response and predictors. Specifically, the response is shuffled relative to the seven predictors. These data are used to test how the OOB error rate of the Random Survival Forest model performs in predicting time to resolution of a missing case when there are no predictors related to the response. The OOB error rate results of the RSF model with logrank and logrankscore splitting criteria as well as the permuted response are shown in Table 2.

Sample size:	3897
Number of events:	3791
Number of trees:	1000
OOB error rate for permuted response and predictors (both rules)	53.2%
OOB error rate for Splitting rule: logrank	46.3%
OOB error rate for splitting rule: logrankscore	47.6%

Table 2: RSF output comparing logrank and logrankscore splitting rules to permuted response using the training data set.

After fitting the RSF with permuted response and predictors, that is shuffling response against its predictors, the OOB error rate is 53.2%. Even though this rate is closer to the random guessing rate which is 50%, this rate is worse than guessing which would produce an error rate of 50%. Using the logrank rule as the splitting criterion, the OOB error rate for the RSF is 46.3%. The error rate increased to 47.6% when the logrankscore splitting criterion is used, suggesting some improvement in prediction versus guessing or versus training on a permuted response, at least using OOB error rates.

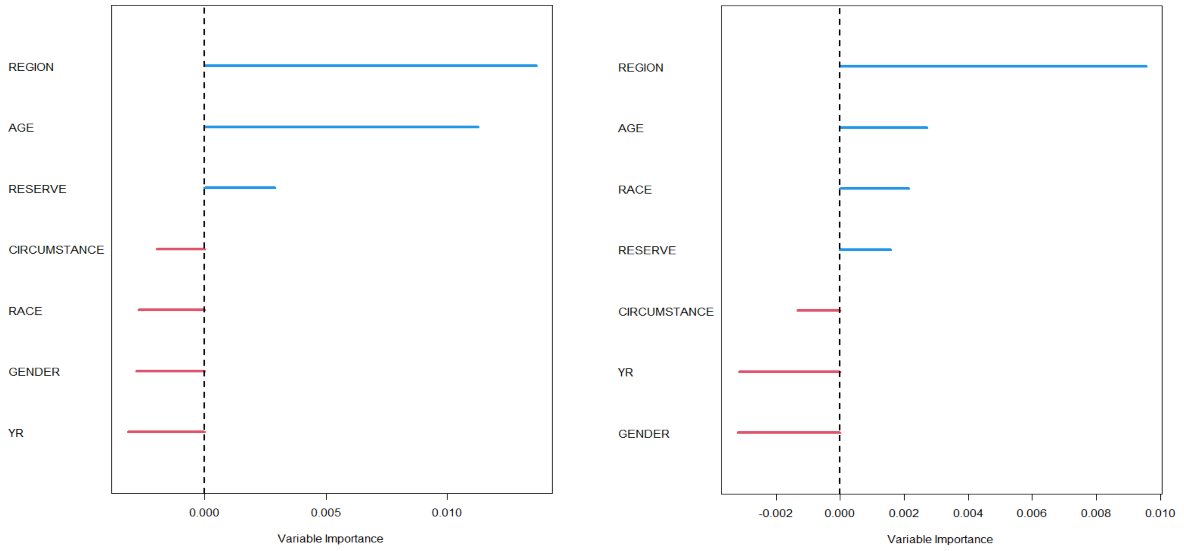
Using the test data to assess the RSF model fitted with the training data, Table 3 shows the test set error rate of the models. The test data set shows similar results for the two splitting criterion as the OOB error rates based just on the training data. In fact, the rates in the test data were slightly lower that the OOB estimates. The error rate for the logrank RSF and logrankscore RSF are 46.1% and 47.2%, respectively. Even though the error rates seem high, the results in Table 3 show that the two splitting rules are better than random guessing.

Sample size:	1669
Number of events:	1630
Number of trees:	1000
Error rate for Splitting rule: logrank	46.1%
Error rate for splitting rule: logrankscore	47.2%

Table 3: Prediction of RSF model on the test data set for the logrank RSF and logrankscore RSF

3.3.1 Variable Importance Results

The variable importance measures compare how well each of the predictor variables are able to predict the time to resolution of a missing case. Using the logrank and logrankscore splitting rules to fit the RSF model, the variable importance of the seven covariates are shown in Figure 8. Using the logrank splitting criterion, Figure 8a shows that region, age, and reservation have the best predictive ability. This means that these covariates are important to the time to resolution in the RSF model. However, circumstance, race, gender, and year variables are less important to the model. In fact, those variables actually do worse in the model then their permuted version with VIMP < 0. Figure 8b shows the variable importance of the 7 covariates in the model with the logrank score split rule. Region, age, race, and reservation are seen to be more important in the model and circumstance, gender, and year are seen to be less important.



(a) VIMP for logrank RSF model.

(b) VIMP for logrankscore RSF model.

Figure 8: Variable Importance (VIMP) for the covariates of time to resolution.

The relative importance of the the seven covariates are also shown in Table 4. In the logrank model, the region was most influential in the prediction with an importance score of 1 and the variable which was the worst in the prediction was the year variable. Also for the logrankscore model, region is seen to occupy the top position with an importance score of 1 and the variable which is least important is year. The age variable in the logrank model has an importance score of 0.8105 while for the logrank score model the relative importance was 0.2973.

	Relative Importance	
Variable	logrank	logrankscore
REGION	1.0000	1.0000
AGE	0.8105	0.2973
RESERVE	0.2076	0.1589
CIRCUMSTANCE	-0.1394	-0.1821
GENDER	-0.1912	-0.3250
RACE	-0.1931	0.2047
YEAR	-0.2283	-0.3373

Table 4: Relative Importance of the 7 covariates of the RSF, sorted by VIMP for logrank RSF.

3.4 Prediction Performance

3.4.1 Concordance Index (C-Index)

The concordance index (C-index) is used to compare the Cox PH model, the Random Survival Forest model using the logrank split rule, and the Random Survival Forest model using the logrankscore split rule. The C-index gives the probability of concordance between the predicted and the observed survival as a function of time. The C-index for each model evaluated at the duration of the missingness period is shown in Figure 9, with higher values suggesting better predictive performance.

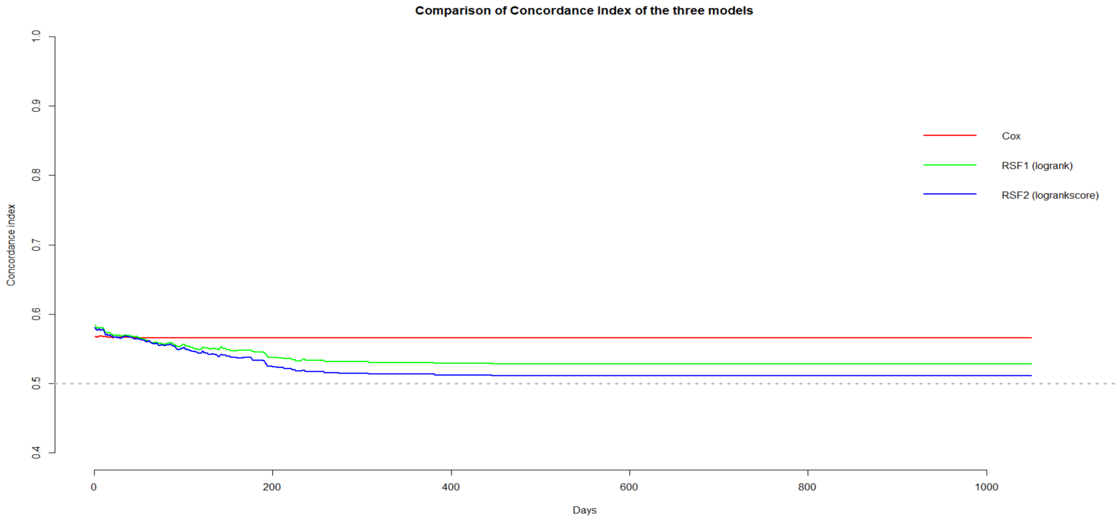


Figure 9: Comparison of Concordance Index (C-Index) over time in days missing.

In Figure 9, it can be seen that the RSF (logrank) model performs better than the remaining two models from 0 days to about 30 days missing in the prediction performance of time to resolution. However, the Cox PH model out performs the other two RSF models after day 30. The RSF (logrankscore) model is the worst performing model on average in the prediction of time to resolution of a missing person case. We can also see in Figure 9 that the C-index for the model drops but at around day 400, they are constant to the the end of the study duration. The reason for the differences in the C-index in the performance of the Cox PH model as compared to the RSF models is unclear, so we also explore Brier scores.

3.4.2 Prediction Error curves using Brier score

The prediction error curves using Brier score is also another way of comparing survival models. Figure 10 shows the prediction error curves for the Kaplan Meier (overall), Cox model, RSF (logrank) model, and RSF (logrankscore) model. This shows the fluctuation of prediction error in each model as time progresses.

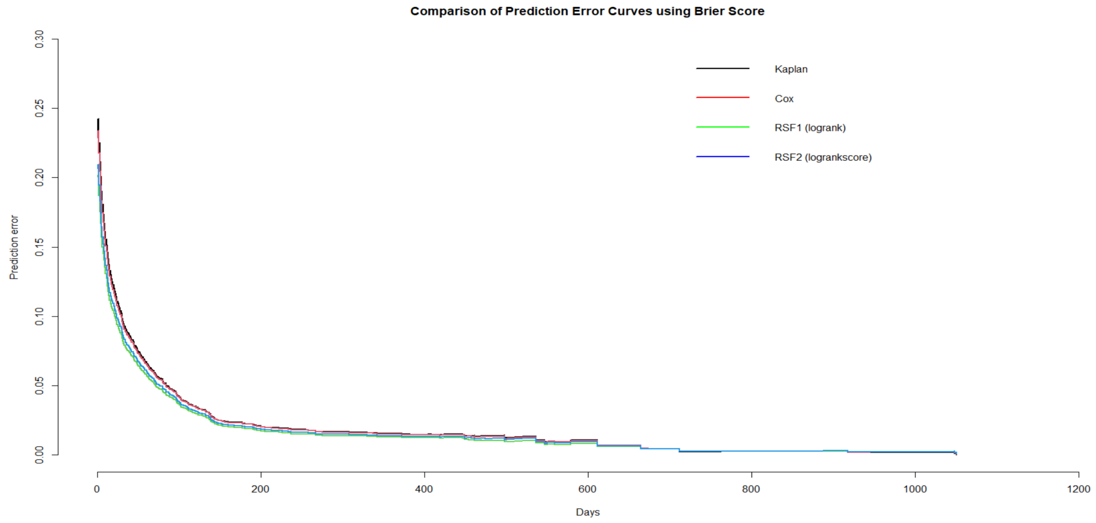


Figure 10: Comparison of prediction error curves from Brier scores.

Looking at Figure 10, it is clear that the two Random Survival Forest models have lower prediction error than the Kaplan-Meier and Cox PH model. The prediction error at day 0 is 0.23 for all the models. This can be attributed to the fact that about 40% of the reported missing case have a zero time to resolution (day 0). The prediction error drops very quickly from 23% to about 3% at day 200. The prediction error curve for the four models levels off after day 200. These results suggest that the Random Survival Forest models performed better than the traditional models.

4 Conclusion and Future work

Random survival forest is a tree-based method for predicting duration times to events given a set of predictors. This paper looked at the application of Random Survival Forest to missing person data in Montana between 2017 and 2019 and compared it with some conventional methods of survival analysis. The Random Survival Forest model used the logrank and logrankscore splitting rule to split candidate variables and we compared the results to a version with a permuted response model. Based on the results, it was seen that the OOB error rate of the RSF model with logrank splitting rule showed the best performance among the three RSF models, even though they are not far from the random guessing threshold.

The circumstance, year, and gender were the least important variables in the models. Race is seen to have a better predictive ability when used in the logrankscore model but is less important when used in the logrank model. The Random Survival Forest models proved to have a lower prediction error than the Cox PH model and the Kaplan Meier model using the Brier score. With the C-index, the Random Survival Forest models performed very well initially but performance dropped as the time progresses. The Cox PH model is seen to have a constant C-Index across time.

A limitation of this study is the structure of the data. The results from this study can only be inferred to missing person cases in Montana in these three years. Only cases reported to authorities are included in these data. There are multiple cases reported on the same person which violates the independence assumption present in all methods used. Missing information on some variables was treated as unknown, which affected the interpretation of some results. Also, the time to resolution was measured in days and so cases that were resolved in hours after reporting were all analyzed as 0 days length events. 40% of the data had time to resolution of 0 days and about 70% of the data had time to resolution from 0 to 3 days. This may have caused the error rates to be high because the subjects with 0 days missing have all been treated as the same with no distinction between events with different hours for predictions. There might be differences in cases missing an hour and those missing more than half a day but these differences can't be modeled or assessed using the data available. Another limitation is that there is suspected interval censoring in the data because the actual start and end times of each case are not known. In our models for these data we only considered right censoring. It is also critically important to note that none of the reported group differences are attributed causally because none of the covariates could be or were randomly assigned, so we only identified associations and the reasons for them is left for future work.

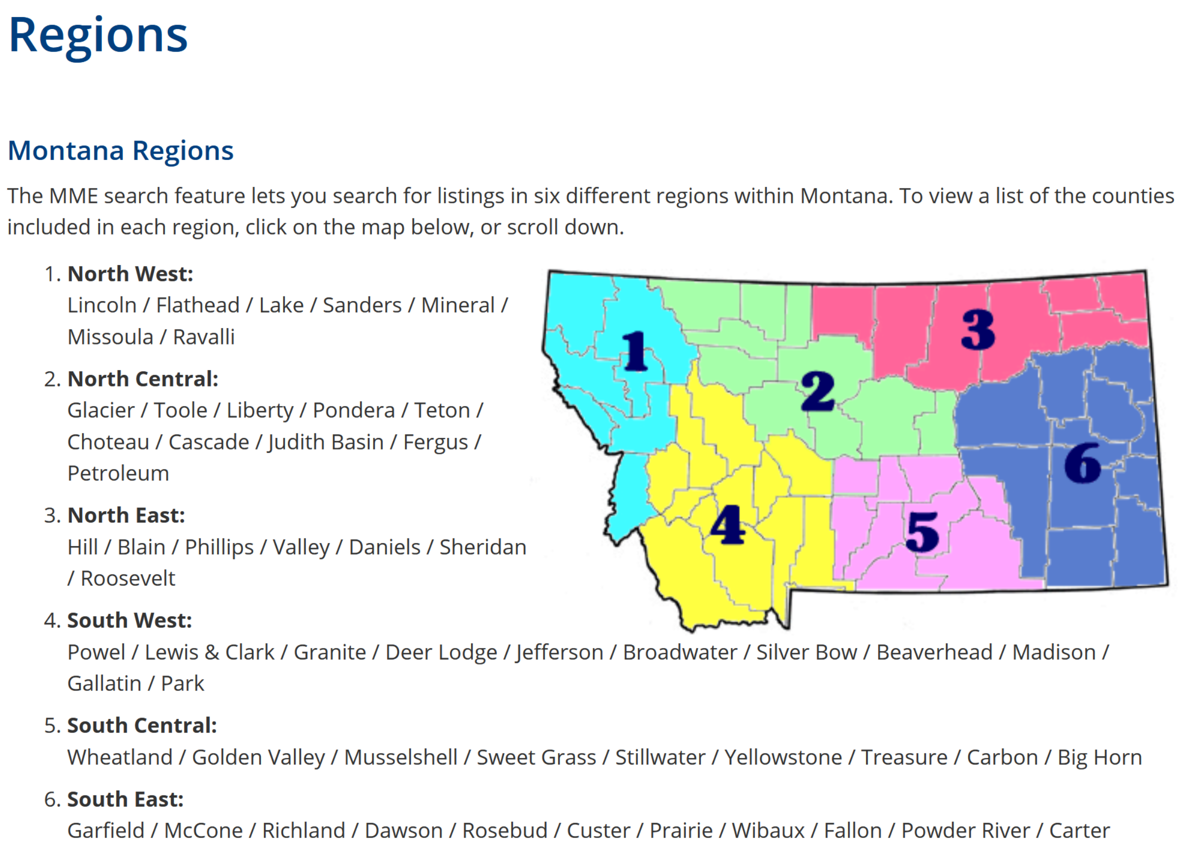
The splitting rule was limited to only the logrank and logrank score splitting rule. Nasejje et al. (2017) noted that there have been some issues regarding the logrank score as it may be linked to the proportional hazards assumption and that can negatively impact the predictive performance of the Random Survival Forest model.

Future work could consider refining the predictor space to remove some of the these less important predictors. Methods such as those discussed in Calle et al. (2011) and Barbour et al. (2019) could be employed to possibly improve the Random Survival Forests. Also, the integrated absolute difference splitting rule discussed in Nasejje et al. (2017) could also be explored. Another Survival Forest model that uses this rule called the Conditional Inference Forest (CIF) could be an ideal alternative to explore for predicting the time to resolution of missing person cases. These models have been known to be a better model than the RSF model because of objectiveness in handling splitting (Nasejje et al., 2017).

The Brier score and Concordance Index are powerful methods of comparing survival models. This study didn't explore deeper into these methods and so this could be an area which could also be explored by looking at how different split methods changes the prediction errors of the survival models.

5 Appendix

5.1 Figures



Regions

Montana Regions

The MME search feature lets you search for listings in six different regions within Montana. To view a list of the counties included in each region, click on the map below, or scroll down.

1. **North West:**
   Lincoln / Flathead / Lake / Sanders / Mineral / Missoula / Ravalli
2. **North Central:**
   Glacier / Toole / Liberty / Pondera / Teton / Choteau / Cascade / Judith Basin / Fergus / Petroleum
3. **North East:**
   Hill / Blain / Phillips / Valley / Daniels / Sheridan / Roosevelt
4. **South West:**
   Powel / Lewis & Clark / Granite / Deer Lodge / Jefferson / Broadwater / Silver Bow / Beaverhead / Madison / Gallatin / Park
5. **South Central:**
   Wheatland / Golden Valley / Musselshell / Sweet Grass / Stillwater / Yellowstone / Treasure / Carbon / Big Horn
6. **South East:**
   Garfield / McCone / Richland / Dawson / Rosebud / Custer / Prairie / Wibaux / Fallon / Powder River / Carter

Figure 11: Montana Materials Exchange (2021). https://www.montana.edu/mme/regions.html

References

Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701 – 726.

Altman, D. G. (1992). Practical statistics for medical research. *Statistics in Medicine*, 10:365–393.

Barbour, C., Kosa, P., Greenwood, M., and Bielekova, B. (2019). Constructing a molecular model of disease severity in multiple sclerosis (p3.2-006). *Neurology*, 92(15 Supplement).

Benyamin, D. (2012). A gentle introduction to random forests, ensembles, and performance metrics in a commercial system. *CitizenNet*, https://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys*, 5(none):44 – 71.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Calle, M. L., Urrea, V., Boulesteix, A.-L., and Malats, N. (2011). Auc-rf: a new strategy for genomic profiling with random forest. *Human Heredity*, 72:121–32.

Camiller, L. (2019). History of survival analysis. https://timesofmalta.com/articles/view/history-of-survival-analysis.705424.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistic Society*, B(34):187–202.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273.

Ishwaran, H. and Kogalur, U. (2007). Random survival forests for °. *R News*, 7(2):25–31.

Ishwaran, H. and Kogalur, U. (2020). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC).* R package version 2.9.3.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, pages 2(3), 841–860.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 1 edition.

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

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, 22(7):263 – 270. Mini-Symposium: Medical Statistics.

Kassambara, A., Kosinski, M., and Biecek, P. (2020). *survminer: Drawing Survival Curves using 'ggplot2'.* R package version 0.4.8.

Koletsi, D. and Pandis, N. (2017). Survival analysis, part 2: Kaplan-meier method and the log-rank test. In *American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics*, volume 152, pages 569–571, United States.

Mishra, S. (2017). Unsupervised learning and data clustering. *Towards data science.*, https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23.

Nasejje, J. B., Mwambi, H., Dheda, K., and Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol*, pages 17, 115.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics : Journal of the International Biometric Society.*, 44(1):35–47.

Therneau, T. M. (2020). *A Package for Survival Analysis in R.* R package version 3.2-7.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5).

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Yiu, T. (2019). Understanding random forest: How the algorithm works and why it is so effective. *Towards Data Science*, https://towardsdatascience.com/understanding-random-forest-58381e0602d2.