FINAL DEGREE PROJECT

Random Survival Forests

Presented by:
Luis Marín Maqueda

Supervised by:
DR. EMILIO CARRIZOSA PRIEGO

FACULTY OF MATHEMATICS
Statistics and Operational Research Department
Seville, May 2021

Contents

Abstract

Survival Analysis is a crucial problem in statistics that tries to explain and model the behaviour of the individuals of a population through some features, one of them the time elapsed until death occurs.

In this work, different models applied in survival analysis are reviewed, from the most traditional estimators to some actual algorithms based in machine learning as the survival trees and forests.

The different models described are compared using the statistical package R with data from the literature.

Introduction

Survival Analysis is a crucial problem in statistics that tries to explain and model the behaviour of the individuals of a population through some features, one of them the time elapsed until death occurs.

In this work, different models applied in survival analysis are reviewed, from the most traditional estimators to some actual algorithms based in machine learning as the survival trees and forests.

This work has the following structure: In chapter 1, the Survival Analysis problem is introduced, concepts as the notion of censoring and some statistic functions used in survival are defined. In chapter 2 two estimators traditionally used in Survival Analysis, namely, Kaplan-Meier estimator and Nelson-Aalen estimator, are defined. In chapter 3 the random tree and the random forests algorithms are discussed. In chapter 4 a measure of the performance of the model obtained, called c-index, and a measure of the weight of the features in the model, called variable importance, are defined. Finally, in chapter 5 we compare the performance of the models proposed using the statistical package R.

Chapter 1

Survival Analysis

Survival Analysis is a statistical problem that consists in explaining and modeling data obtained from measuring different features to some individuals of a population where one of those features is the time elapsed until a certain event occurs.

Unfortunately, measuring the elapsed time until the considered event may be complicated since it cannot be measured instantaneously as it could be with other attributes. Instead, it requires an observation time that can also be interrupted for reasons that are not of interest.

1.1 Censoring

One of the difficulties this analysis presents is censoring, which happens when the time to event is unknown for some of the individuals. Such unknowledge in data can be produced for example when one of the individuals is lost during the observation period for any reason before the event occurred. When data is censored, there is a lack of information, but not all is lost, as it is known that the individual did not reach the event before the time it was observed, so there is still some information.

Definition 1.1.1. Let T be the time elapsed until an individual reach an event of interest. This individual is said to present right-censoring (left-censoring) if all the information given about T is that $T \geq C$ $(T \leq C)$ for a certain $C \in \mathbb{R}$. The individual is said to present interval-censoring if all the information given about T is that $C_1 \leq T \leq C_2$ with $C_1, C_2 \in \mathbb{R}, C_1 \leq C_2$.

In practice, all data is censored as there is no measuring instrument with infinite precision. However, if the associated interval of time is small enough, compared with the observation time, then it can be considered as an uncensored time.

1.2 Survival and hazard functions

In statistics, these functions are related to the cumulative distribution function of a random variable. In particular, in Survival Analysis, they are useful to specify the distribution of the variable T.

Let T be a non negative and absolutely continuous random variable with distribution function $F(t) = P(T \leq t)$. Then, it exits a non-negative function f called the density function such that $F(t) = P(T \leq t) = \int_{-\infty}^{t} f(s)ds$. This is $f(t) = \frac{dF(t)}{dt}$.

Definition 1.2.1. The survival function is defined as the probability of the event is reached beyond the time t, $S(t) = P(T > t) = 1 - F(t)$.

Theorem 1.2.1. *The survival function has the following properties*

1. $S(t)$ *is a monotonically non-increasing function.*
2. $S(t) = 1 \; \forall t \leq 0$
3. $\lim_{t \to \infty} S(t) = 0$

Definition 1.2.2. The hazard function is defined as

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

Theorem 1.2.2. *The hazard function verifies*

$$h(t) = -\frac{d}{dt} \ln(S(t)). \tag{1.1}$$

Proof. We know that

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -\frac{dS(t)}{dt},$$

therefore substituting and operating the result is obtained

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \frac{dS(t)}{dt} = -\frac{d}{dt} \ln(S(t)).$$

□

Definition 1.2.3. The function

$$H(t) = \int_0^t h(s)ds$$

is called the cumulative hazard function.

Theorem 1.2.3. *The cumulative hazard function is related with the survival function by the equation*

$$S(t) = \exp(-H(t)). \tag{1.2}$$

Proof. If the differential equation (1.1) is integrated, then it is obtained $\int_0^t -h(s)ds = \ln(S(t)) - \ln(S(0))$. As it is assumed that $T \geq 0$, then $S(0) = 1$, so $\ln(S(0)) = 0$ and it follows that

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) = \exp(-H(t)).$$

□

1.3 Notation

Let $(T, \underline{X})$ be the random vector of the population of interest. Here, T is the time to the event which will be considered as a non-negative ($T \geq 0$) and absolutely continuous random variable and $\underline{X}$ is a p-variate random vector (that takes values in a sample space $\mathcal{X}$) whose components are the rest of the features considered.

If the sample is incomplete, in the sense that it presents some right-censoring, then there would be $L \in \mathbb{R}$, called the limit of observation. This limit is also a random variable, as it can occur accidentally during the observation time. However, unlike the random variable T, it is not necessarily continuous; if the experiment ends before the individual reached the event, then the probability to reach the limit of observation at the end of the experiment would be nonzero.

In the case where right-censoring is present, the observed time is not necessarily the time T. Instead, it will be observed a vector (Y, Δ) where Y is defined as $Y = \min\{T, L\}$ and Δ is the binary variable defined as

$$\Delta = \begin{cases} 1 & \text{if} \quad T \leq L \\ 0 & \text{if} \quad T > L \end{cases}.$$

When $\Delta = 1$ it is traditionally called a *death* and when $\Delta = 0$ it is called a *loss*.

The random vector $(Y, \Delta, \underline{X})$ has a cumulative distribution function $\bar{F}$, $\bar{F}(y, d, \underline{x}) = P(Y \leq y, \Delta \leq \delta, \underline{X} \leq \underline{x})$.

The objective of Survival Analysis is to infer, from an obtained realization of the simple random sample $(Y_1, \Delta_1, \underline{X}_1), (Y_2, \Delta_2, \underline{X}_2), ..., (Y_n, \Delta_n, \underline{X}_n)$, the distribution of T as a function of the predictor variables $\underline{X}$. This is, for each fixed value $\underline{x}$ of $\underline{X}$, the function $F_{\underline{x}}(t) = P(T \leq t | \underline{X} = \underline{x})$, the probability of reaching the event before the time t of an individual with $\underline{x}$ characteristics.

Theoretically, the observed times are all different for those individuals that reached the event as T is an absolutely continuous variable and the probability of two individuals reaching the event at the same time is null. However, in practice, the values of the time obtained in the realization are not continuous but discrete and ties may occur.

Let $\mathcal{T} = \{y_i : \delta_i = 1\}$ be the set of times to observed event and $q = \#(\mathcal{T})$, the cardinality of such set. Then there is a unique chain $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_q : \mathcal{T} = \{\tau_1, \tau_2, ..., \tau_q\}$. The times τ_i are the different times at which an event is observed $\forall i \in \{1, ..., q\}$, τ_0 is defined just by convention. Moreover n_i is defined as the number of individuals under observation just before time τ_i, this is

$$n_i = \#\{j : t_j \geq \tau_i\} \qquad \forall i \in \{1, ..., q\}.$$

Also d_i is defined as the number of individuals that reach the event (deaths) at time τ_i, mathematically

$$d_i = \#\{j : \delta_j = 1, t_j = \tau_i\} \qquad \forall i \in \{1, ..., q\}.$$

Chapter 2

Traditional estimators

In this section, two of the most traditionally used sample statistics will be introduced because of their importance in Survival Analysis and as a basis for the survival random forest algorithm.

These estimators can be used when only right-censoring is presented. They are both non-parametric estimators, this means that no hypothesis is assumed about the shape of the survival function. Non-parametric estimators are useful when the data does not follow any recognizable distribution or to identify if data follows any known shape so if it does, then a parametric method could be used.

To use these estimators some hypotheses must be assumed:

- The time elapsed until the event occurs T and the observation limit L must be independent variables.
- All data follows the same survival function, there is no dependence with predictor variables.

2.1 Kaplan-Meier estimator

The *Kaplan-Meier estimator*, introduced in [6], is a sample statistic traditionally used in Survival Analysis. This statistic is the maximum likelihood estimator, not of the cumulative distribution, F, but of the survival function defined in 1.2.1.

The Kaplan-Meier estimator is obtained from the product of the estimations of the conditional probabilities $P_i = P(T > \tau_i | T > \tau_{i-1}) = 1 - P(T \leq \tau_i | T > \tau_{i-1})$ $\forall i \in \{1, ..., q\}$.

An estimator of the probability $P(T \leq \tau_i | T > \tau_{i-1})$ is the proportion of individuals that have reached the event at time τ_i among those individuals for which there

is information available at time τ_i that have not reached the event before τ_{i-1}. An estimator of P_i is then

$$\hat{P}_i = 1 - \frac{d_i}{n_i}. \tag{2.1}$$

Theorem 2.1.1. *With the notation above,* $P_i = \dfrac{P(T > \tau_i)}{P(T > \tau_{i-1})}$.

Proof. By the definition of the conditional probability

$$P_i = P(T > \tau_i | T > \tau_{i-1}) = \frac{P((T > \tau_i) \cap (T > \tau_{i-1}))}{P(T > \tau_{i-1})}.$$

As $\tau_i > \tau_{i-1}$, then $P((T > \tau_i) \cap (T > \tau_{i-1})) = P(T > \tau_i)$, so the result is obtained.

$$P_i = \frac{P(T > \tau_i)}{P(T > \tau_{i-1})}.$$

□

Corollary 2.1.1. *The survival function at time* τ_i *satisfies* $S(\tau_i) = \prod_{j=1}^{i} P_j$.

Proof. From the definition of the survival function it is obtained that

$$S(\tau_i) = P(T > \tau_i) = P(T > \tau_0) \prod_{j=1}^{i} \frac{P(T > \tau_j)}{P(T > \tau_{j-1})}.$$

As T is a continuous and positive variable, we have that $P(T > \tau_0 = 0) = 1$. Then, using the theorem above the result is obtained.

$$S(\tau_i) = \prod_{j=1}^{i} \frac{P(T > \tau_j)}{P(T > \tau_{j-1})} = \prod_{j=1}^{i} P_j.$$

□

The Kaplan-Meier estimator for the survival function is then defined as

$$\hat{S}_{KM}(t) = \prod_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \hat{P}_i = \prod_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \left(1 - \frac{d_i}{n_i}\right) \tag{2.2}$$

2.2 Nelson-Aalen estimator

The *Nelson-Aalen estimator* is another sample statistic used in Survival Analysis. It was introduced by Nelson in [9] and by Aalen in [1]. This statistic is an estimator of the cumulative hazard function.

The Nelson-Aalen estimator of the hazard function is defined as

$$\hat{H}_{NA}(t) = \sum_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \frac{d_i}{n_i} \tag{2.3}$$

This estimator can be transformed as in equation (1.2) to obtain an estimator of the survival function, namely

$$\hat{S}_{NA}(t) = \exp\left(-\hat{H}_{NA}(t)\right) = \exp\left(-\sum_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \frac{d_i}{n_i}\right) = \prod_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \exp\left(-\frac{d_i}{n_i}\right) \tag{2.4}$$

Theorem 2.2.1. *The Kaplan-Meier and the Nelson-Alen estimator for the survival funtion satisfy*

$$\hat{S}_{KM}(t) \leq \hat{S}_{NA}(t) \quad \forall t \in \mathbb{R} \tag{2.5}$$

Proof. Let us fix $t \in \mathbb{R}$, then

$$\frac{\hat{S}_{KM}(t)}{\hat{S}_{NA}(t)} = \frac{\prod\limits_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \left(1 - \frac{d_i}{n_i}\right)}{\prod\limits_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \exp\left(-\frac{d_i}{n_i}\right)} = \prod_{\substack{i:\tau_i \leq t \\ i \in \{1,\dots,q\}}} \frac{\left(1 - \frac{d_i}{n_i}\right)}{\exp\left(-\frac{d_i}{n_i}\right)}$$

It is known that $\frac{1-x}{e^{-x}} \leq 1$ so for each $i \in \{1, ..., q\} : \tau_i \leq t$ it is $\dfrac{\left(1 - \frac{d_i}{n_i}\right)}{\exp\left(-\frac{d_i}{n_i}\right)} \leq 1$ and the result is obtained, $\frac{\hat{S}_{KM}(t)}{\hat{S}_{NA}(t)} \leq 1$. □

An illustrative example will be provided to show how do these two traditional statistics work.

Example. Five individuals are observed for a week and the time elapsed before they reach the event of interest is measured. The following information is obtained:

Individual i	Observed time y_i (days)	Censoring δ_i
1	3	0
2	1	1
3	5	1
4	7	0
5	4	0

Table 2.1: Information obtained in an illustrative example

With this data, we have that $q = 2$ and $(\tau_1, \tau_2) = (1, 5)$, then $(n_1, n_2) = (5, 2)$ and $(d_1, d_2) = (1, 1)$.

The quotients $\frac{d_i}{n_i}$ in this example are $\frac{d_1}{n_1} = \frac{1}{5}$ and $\frac{d_2}{n_2} = \frac{1}{2}$. The estimations for the conditional probabilities P_i are calculated with the formula (2.1) so $\hat{P}_1 = 1 - \frac{1}{5} = \frac{4}{5}$ and $\hat{P}_2 = 1 - \frac{1}{2} = \frac{1}{2}$.

Then the Kaplan-Meier estimator of the survival function is obtained using the equation (2.2)

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{if} \quad t < 1 \\ 0.8 & \text{if} \quad 1 \leq t < 5 \\ 0.4 & \text{if} \quad t \geq 5 \end{cases}$$

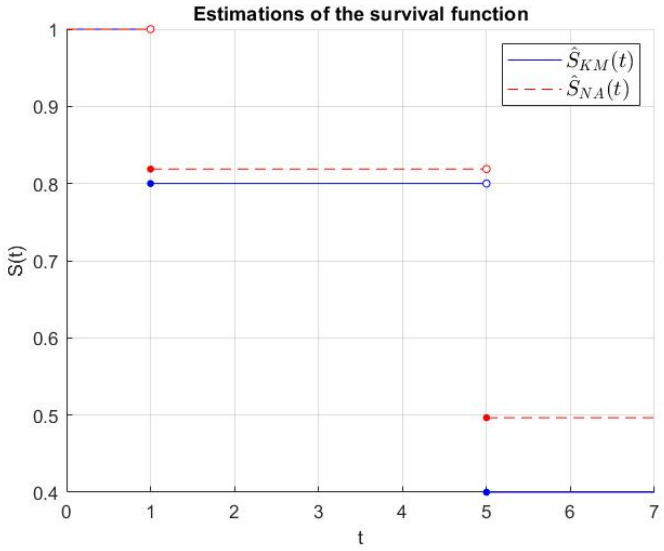
The Nelson-Aalen estimator of the hazard function is obtained with the equation (2.3)

$$\hat{H}_{NA}(t) = \begin{cases} 0 & \text{if} \quad t < 1 \\ 0.2 & \text{if} \quad 1 \leq t < 5 \\ 0.7 & \text{if} \quad t \geq 5 \end{cases}$$

The transformation of the Nelson-Aalen estimator as an estimator of the survival function given in (2.4) is:

$$\hat{S}_{NA}(t) = \begin{cases} 1 & \text{if} \quad t < 1 \\ 0.819 & \text{if} \quad 1 \leq t < 5 \\ 0.497 & \text{if} \quad t \geq 5 \end{cases}$$

In the next image, both estimations of the survival function are represented.



In this example it is seen that the estimations for the survival function provided by the Kaplan-Meier estimator and the Nelson-Aalen estimator are not survival functions as $\lim_{t\to\infty} \hat{S}(t) > 0$. This also implies that no mean or variance can be directly estimated from the estimated survival function. The unique case the estimated function is a survival function is for the Kaplan-Meier estimator when the last observation is a death as then $d_q = n_q$ and $\left(1 - \frac{d_q}{n_q}\right) = 0$ so $\hat{S}_{KM}(t) = 0 \; \forall t \geq \tau_q$.

This problem is often solved considering a time $t^* \geq \tau_q$ and defining

$$\hat{S}'(t) = \begin{cases} \hat{S}(t) & \text{if} \quad t < t^* \\ 0 & \text{if} \quad t \geq t^* \end{cases}$$

which is a survival function and mean and variance can be defined.

Chapter 3

Random Forests

Random forests are a machine learning technique that was introduced by Breiman in [2]. While they were first proposed for regression and classification, they can easily be modified to be applied in Survival Analysis [4].

To understand what a random forest is, it is necessary to introduce the concept of a random tree first.

3.1 Random Trees

A random tree is a tree structure with an association between each node in the tree and the subsets of a sample space $\mathcal{X}$ in a hierarchical way and a prediction at the endings. Below will be explained the tree structure, the association between the subsets of the sample space and the tree structure, and the prediction at the endings of the random tree.

3.1.1 Tree structure

The kind of trees considered in this text consist of a finite set of nodes $\mathcal{T} = \{N_1, N_2, ..., N_m\}$ hierarchically related in the way explained below.

- The node N_1 is called the *root* node.
- Each node N_i with $i \geq 2$ is related to a single node N_j for some $j < i$. The node N_j is said to be the *mother* of the node N_i.
- Each node N_i verifies one and only one of these statements:
  - Either it is not related to any other node than its mother, in which case N_i is called a *leaf*.

  - Either N_i is related to exactly two nodes N_j and N_k with $i < j, k$ that are its *daughters*.

- A node N_i is the mother of a node N_j if and only if N_j is one of the daughters of N_i.

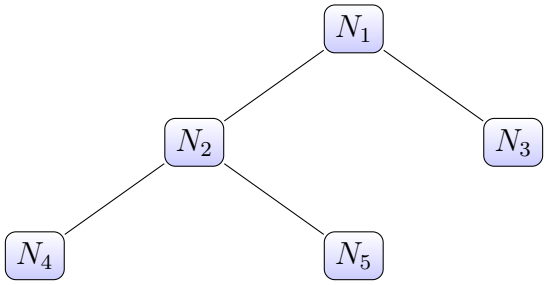


Figure 3.1

The figure 3.1 represents an example of tree. The root is the node N_1 and the leaves are the nodes N_3, N_4 and N_5. The node N_1 is the mother of the nodes N_2 and N_3 and the node N_2 is the mother of the nodes N_4 and N_5. Equivalently, the nodes N_2 and N_3 are the daughters of the node N_1 and the nodes N_4 and N_5 are the daughters of the node N_2.

3.1.2 Association between subsets of the sample space and nodes

With each node N of the tree we associate a subset $\mathcal{X}_N$ of the sample space $\mathcal{X}$ satisfying the following

- The sample space $\mathcal{X}$ is associated with the root node N_1, $\mathcal{X}_{N_1} = \mathcal{X}$.

- If the node N_i is the mother of the nodes N_j and N_k then $\mathcal{X}_{N_i} = \mathcal{X}_{N_j} \cup \mathcal{X}_{N_k}$ and $\mathcal{X}_{N_j} \cap \mathcal{X}_{N_k} = \varnothing$.

For simplicity in the notation, we identify each node N with its associated subset $\mathcal{X}_N$.

An illustrative example is shown in the figure 3.2 with the tree structure in figure 3.1 and the finite sample space $\mathcal{X} = \{a, b, c, d, e, f, g, h, i, j\}$

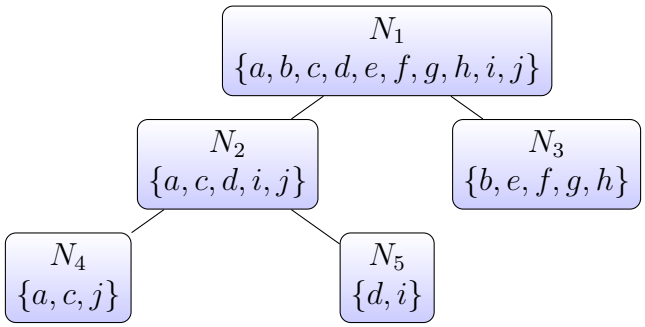


Figure 3.2

3.1.3 Prediction at leaves

The random tree $\mathcal{T}$ defines a partition into disjoint subsets of the sample space $\mathcal{X}$ at the leaf nodes. Let $\mathcal{L}$ be the set of the leaves in $\mathcal{T}$, then by definition $\bigcup_{L \in \mathcal{L}} \mathcal{X}_L = \mathcal{X}$ and $\mathcal{X}_{L_1} \cap \mathcal{X}_{L_2} = \varnothing \quad \forall L_1, L_2 \in \mathcal{L} : L_1 \neq L_2$.

If the behaviour is roughly the same for all the individuals in the same leaf, then it would be reasonable to try to find a prediction for each leaf instead of for the whole sample space. This prediction could be the mean for a classification tree, a regression curve for a regression tree or, what is of most interest in this work, an estimator of the survival function for a survival tree.

3.2 Growing a random survival tree

Given a sample realization $(y_1, \delta_1, \underline{x}_1), (y_2, \delta_2, \underline{x}_2), ..., (y_n, \delta_n, \underline{x}_n)$ from a simple random sample of the random vector $(Y, \Delta, \underline{X})$, the goal is to grow a random tree so at each leaf node all the individuals are expected to have the same behaviour. In the Survival Analysis case, they are expected to have the same survival function, which can be estimated through the Kaplan-Meier estimator or the Nelson-Aalen estimator at each leaf node.

In an algorithm that grows a random tree, it is necessary to specify a criterion to split the data associated in each node into two disjoint subsets and also a stopping criterion, this is, when should the algorithm consider that a node is a leaf node.

Given a sample $\mathcal{D}$, the number of ways the finite set $\mathcal{D}$ can be split into two disjoint subsets is $2^{\#\mathcal{D}-1} - 1$ (the subsets $\varnothing$ and $\mathcal{D}$ are not considered as a split). This means that it increases exponentially fast, so it is unthinkable to try to select a split out of all the possible ones even for a set of moderate cardinality. Furthermore, if all the possible

splits were considered, the chosen split can be difficult to interpret.

It is then necessary to reduce the set of splits permitted. A way to reduce the number of permitted splits is to consider only the simplest and the easiest to interpret, these are the splits that divide the sample according to only one single predictor variable. This is, fixed the predictor variable X^j of the vector $\underline{X}$, the type of splits allowed are:

- if X^j is an ordered variable, then the split consists of forming the disjoint subsets $\{i : x_i^j \leq c\}$ and its complementary $\{i : x_i^j > c\}$ where $c \in \mathbb{R}$ can be chosen.

- if X^j is a nominal variable that takes values in the finite set $\mathcal{W} = \{w_1, ..., w_{k_j}\}$, then the subsets that provide the split are $\{i : x_i^j \in S\}$ and its complementary $\{i : x_i^j \notin S\}$ where $S \subset \mathcal{W}$.

The constant c defined before can apparently take uncountable values, producing infinite different splits, nevertheless, given the sample $\mathcal{D}$, the variable X^j will at most take $\#\mathcal{D}$ different values, so it could be enough to consider $\#\mathcal{D}-1$ values of c between every two consecutive values of the variable X^j to consider all the possible splits in that variable.

The criteria used to choose the best split among those allowed at each node in a random tree normally consists of considering a measure of dissimilarity or a distance between two sets. The split chosen is the one that maximizes the distance between the disjoint subsets associated to the daughter nodes.

The criterion to select a split out of all the considered splits depends on the kind of analysis for which the random tree is grown. Some criteria used to split random trees applied in Survival Analysis can be found in section 4.1 in [13]

One of those criteria is the log-rank criterion, discussed in [8] and [11]. This criterion is among a family of estimators called the Tarone-Ware class of statistics and is described below.

Let us consider an allowed split at a node in the random tree that is been grown. This split divides the data set $\mathcal{A}$ associated to the node into two different populations $\mathcal{B}$ and $\mathcal{C}$. Let us call $q_{\mathcal{A}}$ the number of different times at which the event is observed in any of the individuals of the population $\mathcal{A}$. For each $i \in \{1, ..., q_{\mathcal{A}}\}$, $\tau_{\mathcal{A},i}$ is the i-th observed time to event in $\mathcal{A}$. We define $d_{\mathcal{A},i}$ as the number of deaths in $\mathcal{A}$ at time $\tau_{\mathcal{A},i}$ and $d_{\mathcal{A},i,\mathcal{B}}$ as the number of deaths in $\mathcal{B}$ at time $\tau_{\mathcal{A},i}$. They are also defined $n_{\mathcal{A},i}$ and $n_{\mathcal{A},i,\mathcal{B}}$ as the number of individuals under observation before time $\tau_{\mathcal{A},i}$ in $\mathcal{A}$ and $\mathcal{B}$ respectively.

	Dead	Alive	Sum
Population $\mathcal{B}$	$d_{\mathcal{A},i,\mathcal{B}}$		$n_{\mathcal{A},i,\mathcal{B}}$
Population $\mathcal{C}$			
Population $\mathcal{A}$	$d_{\mathcal{A},i}$		$n_{\mathcal{A},i}$

Table 3.1: An illustrative table at time $\tau_{\mathcal{A},i}$ to better understand these definitions

The class of Tarone-Ware statistics has the form:

$$TW_{\mathcal{A},\mathcal{B}} = \frac{\sum_{i=1}^{q_{\mathcal{A}}} w_i \left(d_{\mathcal{A},i,\mathcal{B}} - \hat{E}_{0,i,\mathcal{B}} \right)}{\sqrt{\sum_{i=1}^{q_{\mathcal{A}}} w_i^2 \hat{V}_{0,i,\mathcal{B}}}} \tag{3.1}$$

where $\hat{E}_{0,i,\mathcal{B}}$ and $\hat{V}_{0,i,\mathcal{B}}$ are respectively the expectation and the variance of the number of deaths in the population $\mathcal{B}$ at time $\tau_{\mathcal{A},i}$ under the null hypothesis: the death rates are the same at each instant $\tau_{\mathcal{A},i}$ in both populations $\mathcal{B}$ and $\mathcal{C}$. The null hypothesis assumption implies that the number of deaths in the population $\mathcal{B}$ follows a hypergeometric distribution so

$$\hat{E}_{0,i,\mathcal{B}} = \frac{d_{\mathcal{A},i} n_{\mathcal{A},i,\mathcal{B}}}{n_{\mathcal{A},i}} \qquad \text{and}$$
$$\hat{V}_{0,i,\mathcal{B}} = \left[\frac{d_{\mathcal{A},i}(n_{\mathcal{A},i} - d_{\mathcal{A},i})}{n_{\mathcal{A},i} - 1} \right] \left[\frac{n_{\mathcal{A},i,\mathcal{B}}}{n_{\mathcal{A},i}} \left(1 - \frac{n_{\mathcal{A},i,\mathcal{B}}}{n_{\mathcal{A},i}} \right) \right].$$

The constants w_i are weights associated to each table. The most simple and known statistic from the Tarone-Ware class is the log-rank statistic which is obtained with the weights $w_i = 1$. Other statistics employed are the Gehan statistic, obtained with $w_i = n_{\mathcal{A},i}$; the Tarone-Ware statistic, obtained with $w_i = \sqrt{n_{\mathcal{A},i}}$ and the Harrington-Fleming statistic, when $w_i = \hat{S}_{KM}(\tau_{\mathcal{A},i})$ is the Kaplan-Meier survival estimation at time $\tau_{\mathcal{A},i}$. For more information see [11].

The chosen split for each node is the one (among all the allowed splits at the node) that maximizes the absolute value of one of the Tarone-Ware statistics, for example, the log-rank statistic.

There are many criteria to decide when a node is considered as a leaf so it has no daughters but all of them basically entail that the data associated to the nodes is not too small. In Survival Analysis with right-censored data, a necessary criterion is that for each leaf node at least one of the individuals in the associated data must reach the event. This is, all the individuals cannot be censored as otherwise, no estimations are possible. A commonly used criterion is to fix $d_0 > 0$ and impose that the number of deaths at each node is bigger or equal than d_0.

Consider the data set $(y_1, \delta_1, \underline{x}_1), (y_2, \delta_2, \underline{x}_2), ..., (y_n, \delta_n, \underline{x}_n)$ obtained from a simple random sample of $(Y, \Delta, \underline{X})$ which takes values in $(0, +\infty) \times \{0, 1\} \times \mathcal{X}$. The algorithm to grow a random survival tree is described below

Algorithm

1. All the individuals are associated with the root.
2. At each node repeat
   (a) Randomly select r from the p features of $\underline{X}$.
   (b) Consider all the splits in those r variables that guarantee that the daughter nodes satisfy that there are at least $d_0 > 0$ deaths associated to each one.
   (c) If no split has been considered, then this node is a leaf and the algorithm continues with another node.
   (d) If some splits have been considered, then choose the one that maximizes the log-rank statistic and grow the tree adding two daughter nodes.
3. At each leaf node an estimator of the survival function is computed from the individuals in the node, for example, the Kaplan-Meier estimator.

Usually, growing one single random tree does not provide a satisfactory model as it can provide very different models just because of random selections of the parameters at each node. This problem can be solved by growing some trees from the same data and choosing the one that provides the best model. However, better results can be obtained if the trees are considered together in a random forest.

3.3 Random survival forests

A random forest consists of a set of random trees that together are expected to provide a stronger and more sophisticated model than a unique tree. When a random forest is applied to Survival Analysis it is called a *random survival forest*.

Again, let us consider the data set $(y_1, \delta_1, \underline{x}_1), (y_2, \delta_2, \underline{x}_2), ..., (y_n, \delta_n, \underline{x}_n)$ obtained from a simple random sample of the random vector $(Y, \Delta, \underline{X})$ of the population of interest. The random survival forest algorithm is described below:

Algorithm

1. B bootstrap samples are randomly chosen. This is, B independent random samples with replacement of the same size as the data are generated. Note that for each bootstrap sample approximately 37% of data is not considered. For each bootstrap sample the not considered data is called OOB (out-of-bag).

2. From each bootstrap sample, for $b \in \{1, ..., B\}$, a random tree $\mathcal{T}_b$ is grown.

3. For each tree $\mathcal{T}_b$ consider all the individuals in the sample at each leaf $L_{j,b} \in \mathcal{L}_b$, this is $\{(y_i, \delta_i, \underline{x}_i) : \underline{x}_i \in \mathcal{X}_{L_{j,b}}\}$, and compute an estimator for the survival function $\hat{S}_{j,b}(t)$, for example the Kaplan-Meier estimator (or the Nelson-Aalen estimator).

4. The estimation of the survival function for the b-th tree is

$$\hat{S}_b(t|\underline{x}) = \sum_{j:L_{j,b}\in\mathcal{L}_b} I_{\mathcal{X}_{L_{j,b}}}(\underline{x})\hat{S}_{j,b}(t)$$

   where $I_{\mathcal{Y}}$ is the indicator function of the subset $\mathcal{Y}$ of the set $\mathcal{X}$.

5. The ensemble survival function is obtained by taking the average of the survival functions of each tree in the forest, this is,

$$\hat{S}_e(t|\underline{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_b(t|\underline{x})$$

6. To predict error also an OOB survival function estimation can be defined. For each $\underline{x}_i$ in the original sample, the OOB survival function is defined as the average of the survival functions of each tree for which $\underline{x}_i$ is not in the bootstrap sample, or equivalently,

$$\hat{S}_{OOB}(t, \underline{x_i}) = \frac{\sum_{b=1}^{B} I_{i,b}\hat{S}_b(t|\underline{x}_i)}{\sum_{b=1}^{B} I_{i,b}}$$

   where $I_{i,b} = 1$ if $\underline{x}_i$ is not in the b-th bootstrap sample and $I_{i,b} = 0$ if $\underline{x}_i$ is in the b-th bootstrap sample.

Similarly, in addition to the ensemble and the OOB survival functions, ensemble and OOB cumulative hazard functions can be calculated, $\hat{H}_e(t, \underline{x})$ and $\hat{H}_{OOB}(t, \underline{x})$ respectively.

Chapter 4

Performance of model and features

4.1 C-index

In order to measure the accuracy of a survival model, an especial estimator is necessary so the censoring is taken into account. The concordance index, also called c-index, first introduced in [3], is defined as the probability that, given a pair of individuals, the prediction model provides a worse outcome for the individual that dies first. This c-index has a similar interpretation to the area under the ROC curve and can handle the censored data so it can be used to estimate the performance of a survival model.

Let us consider that a prediction for the ensemble cumulative hazard function $\hat{H}_e(t|\underline{x})$ has been obtained from the dataset $(y_1, \delta_1, \underline{x}_1), (y_2, \delta_2, \underline{x}_2), ..., (y_n, \delta_n, \underline{x}_n)$. We define the mortality of an individual with features $\underline{x}$ as

$$\mathfrak{M}(\underline{x}) = \sum_{l=1}^{n} \hat{H}_e(y_l|\underline{x}).$$

Mortality does not depend of the instant of time and provides a measure of the outcome so comparisons can be made. An individual with features $\bar{\underline{x}}_1$ is said to have a worst outcome than an individual with features $\bar{\underline{x}}_2$ if $\mathfrak{M}(\bar{\underline{x}}_1) > \mathfrak{M}(\bar{\underline{x}}_2)$.

Let $\mathcal{D} = \left\{ \left(\tilde{y}_1, \tilde{\delta}_1, \tilde{\underline{x}}_1 \right), \left(\tilde{y}_2, \tilde{\delta}_2, \tilde{\underline{x}}_2 \right), ..., \left(\tilde{y}_m, \tilde{\delta}_m, \tilde{\underline{x}}_m \right) \right\}$ be a dataset (which can be the same training data used to obtain $\hat{H}_e$, the OOB data or some new data). Then, considering that $\tilde{y}_i \neq \tilde{y}_j \ \forall i, j : \tilde{\delta}_i = \tilde{\delta}_j = 1$, an estimation of the c-index is

$$C_{\mathcal{D}} = \frac{1}{M} \sum_{i:\tilde{\delta}_i=1} \sum_{j:\tilde{y}_i \leq \tilde{y}_j} \left[\mathbf{I}\left(\mathfrak{M}(\tilde{\underline{x}}_i) > \mathfrak{M}(\tilde{\underline{x}}_j) \right) + 0.5 \cdot \mathbf{I}\left(\mathfrak{M}(\tilde{\underline{x}}_i) = \mathfrak{M}(\tilde{\underline{x}}_j) \right) \right] \tag{4.1}$$

where M is the number of addends in equation (4.1) and $\mathbf{I}(\cdot)$ is the indicator function (it takes the value 1 if the argument is true and 0 otherwise), more information about

this estimation of the c-index can be found in [4]. Another estimator for the c-index is described in [12] and in [13].

With this definition of c-index, if $C_{\mathcal{D}} = 1$, then the concordance of the model is perfect, this is, the model assigns better outcomes, lower mortality, to the individuals that live longer. If $C_{\mathcal{D}} = 0.5$, then the model is not better than randomly choosing which individual will live longer. Finally, if $C_{\mathcal{D}} = 0$, then the model always fails in the prediction of which individual will live longer.

4.2 Variable Importance

In addition to providing a prediction model for the considered population, random trees and random forests can provide an estimation of the importance of a single variable X^j of $\underline{X}$. The variable importance is a measure of the significance of a variable in the random tree or the random forest. In this work, two ways to estimate the variable importance are explained.

The first method, called permutation method, is described in [2]. It consists of randomly permuting the values that the variable X^j takes between the out-of-bag individuals and then to compare the c-index estimated with this permuted data and the c-index obtained with the original *oob* data. Mathematically, this is, once the ensemble survival function $\hat{S}(t|\underline{x})$ and the mortality $\mathfrak{M}(\underline{x})$ have been computed from some training data with OOB data (or any new sample obtained from the the population of interest) $\mathcal{D} = \{(y_1, \delta_1, \underline{x}_1), (y_2, \delta_2, \underline{x}_2), ..., (y_n, \delta_n, \underline{x}_n)\}$, a permutation $\sigma \in S_n$ is randomly chosen. We define $\underline{\tilde{x}}_i = (x_i^1, x_i^2, ..., x_{\sigma(i)}^j, ..., x_i^p)$ and the permuted data $\mathcal{D}'$ as

$$\mathcal{D}' = \{(y_1, \delta_1, \underline{\tilde{x}}_1), (y_2, \delta_2, \underline{\tilde{x}}_2), ..., (y_n, \delta_n, \underline{\tilde{x}}_n)\}\,.$$

The estimation of the importance of the variable X^j is VIMP $= C_{\mathcal{D}} - C_{\mathcal{D}'}$, where $C_{\mathcal{D}}$ and $C_{\mathcal{D}}$ are the concordance indices for the correspondent data.

The other method that will be explained in this work, called random method, is described in [4]. This method consists of dropping the OOB data of each tree down the respective tree but at each node split in the variable X^j each individual is drop randomly to one or another daughter, and then to compare the c-index estimated with the original data and the c-index obtained this way averaged over the trees in the survival forest.

The variable importance, VIMP, has the following interpretation: if VIMP $= 0$, then it means that the prediction does not really depend on the variable X^j, and if VIMP > 0, then the variable helps to explain the survival of the individuals in the model. It could also be that VIMP < 0, in this case, the variable in the model would explain the survival even worse than if the variable was not considered.

Chapter 5

Examples using R

Here we compute some examples to illustrate what has been explained in this work. For this purpose, we have used the R-package `randomForestSRC`.

The main function in the R-package `randomForestSRC` is the function `rfsrc` which computes a Breiman's random forest for different kinds of data such as regression, classification, right-censored survival and competing risk. This function admits several arguments, so only the ones used in this work are described below.

Usage: `rfsrc(formula, data, ntree, nodesize, samptype)`

Arguments

`formula`	A symbolic description of the model to be fit.
`data`	Data frame containing the y-outcome and x-variables.
`ntree`	Number of trees.
`nodesize`	Forest average number of individuals in terminal nodes.
`samptype`	Type of sampling which can be with or without replacement.

The function `rfsrc` provides an object of class (`rfsrc`,`grow`), to obtain information from this object we have used the following commands.

Auxiliary commands

`print`	Provides a summary of the computed forest.
`subsample`	Randomly choose some subsamples from data and computes the variable importance from each subsample obtaining a kind of confidence importance for VIMP.

Example. We will first study the data set Veteran's Administration Lung Cancer Trial presented in [10] which is implemented in the R-package `randomForestSRC` under

the name `veteran`.

This data set consists of 137 advanced lung cancer patients of which 9 are right-censored. The variables recorded for each patient are:

1. the given chemotherapeutic agent (standard treatment, 1; test treatment, 2),
2. the tumour cell type (squamous, 1; small, 2; adeno, 3; and large, 4),
3. a measure of the medical status provided by the Karnofsky scale which is a score between 0 and 100 that assigns higher scores to patients that feel better,
4. the time from diagnosis to the beginning of the study in months,
5. the age of the patient in years and
6. a binary variable taking the value 1 if the patient had got previous therapy and 0 if had not.

We show in figure 5.1 the dependence of the time to event in the sample with the predictor variables. It can be appreciated that patients which celltype is 1 and 4 live longer than those which celltype is 2 or 3 and that patients with high karno scores also live longer than those with lower scores. These observations will be reflected in the variable importance of these two variables.

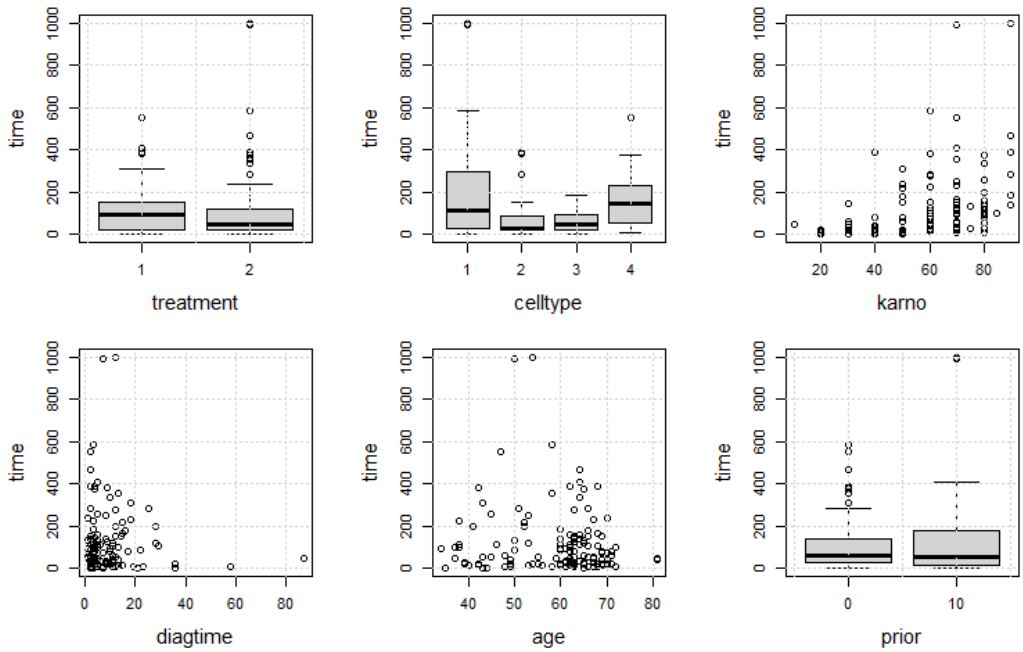


Figure 5.1: Representation of the time for non-censored data versus the predictor variables

In this example we will compute over the data set the Kaplan-Meier estimator, a single random survival tree and the random survival forest algorithm to show and compare the performance of these algorithms.

We have computed the Kaplan-Meier estimator from a subsample without replacement of 87 of the patients using the function `rfsrc` with one single tree and the node size large enough so the tree consists of one single node.

The Kaplan-Meier estimator provides the same survival function for every patient independently of the predictor variables. This implies that all of them have the same estimated mortality so the c-index of the model provided applying the Kaplan-Meier estimator directly is 0.5. In addition, the estimated variable importance is null for each variable as in fact the model does not consider the variables.

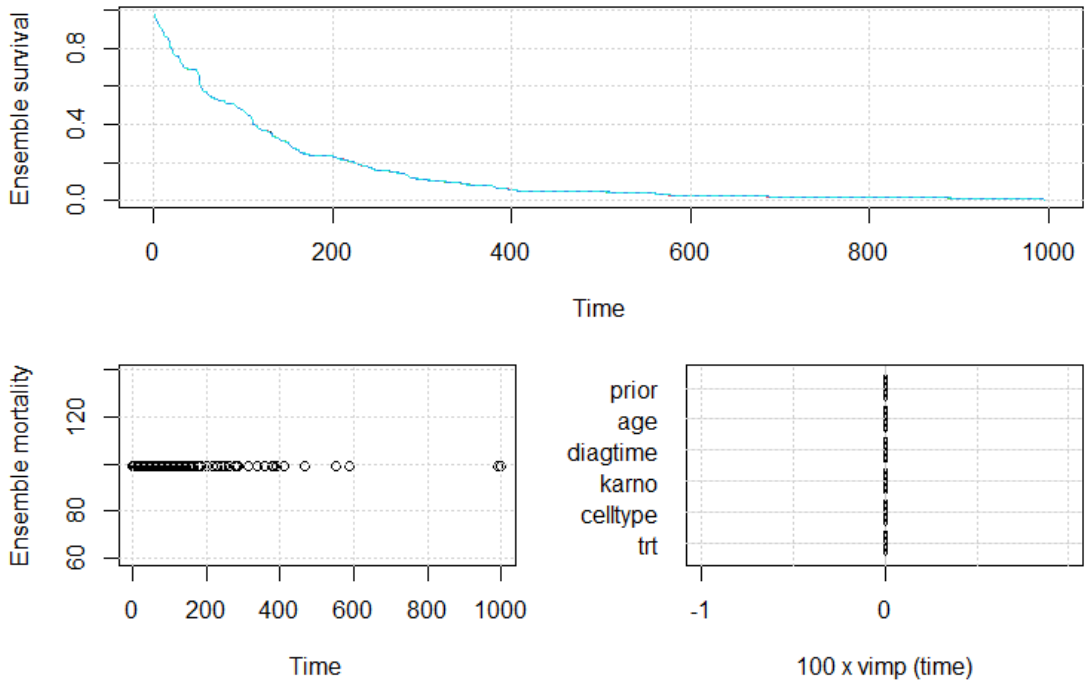


Figure 5.2: Representations of the survival function, the mortality and the VIMP computing the Kaplan-Meier estimator

We have also computed a single random survival tree from a subsample without replacement of 87 of the patients using the function `rfsrc` with one single tree and the node size of 15.

The model obtained divides the patients into 5 different leaves and assigns the

same survival function for all the patients associated to the same leaf. The survival function for each leaf is estimated from the few patients associated to that leaf. The c-index of the model, computed using the out-of-sample data, is near 0.6. The variable importance is computed for each variable and it is seen that karno, age and cell type are important variables in the model. This does not necessarily match with the reality as we have computed only one tree and the variable importance depends on the splits randomly considered at the nodes of the tree.

trt	celltype	karno	diagtime	age	prior
0	0.026	0.077	0	0.044	0

Table 5.1: VIMP in a single tree

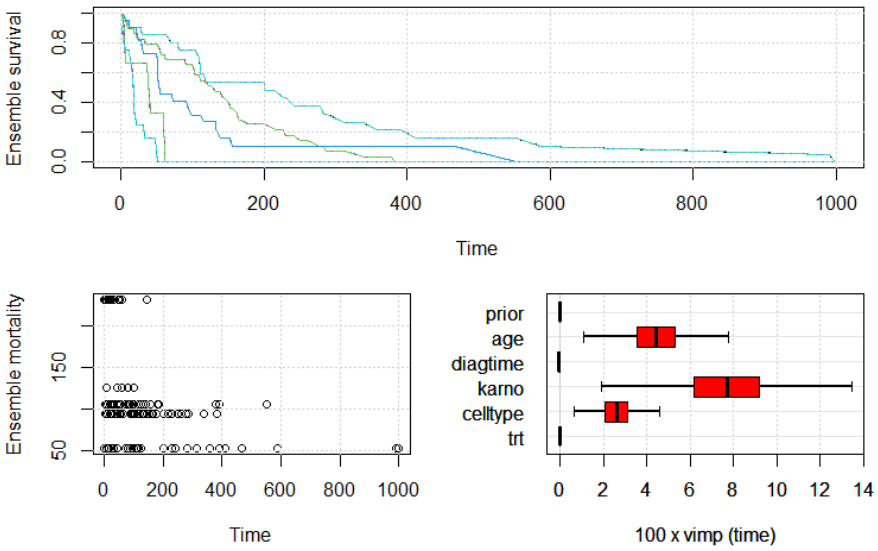


Figure 5.3: Representations of the survival function, the mortality and the VIMP computing a single tree

Finally, we have computed a random survival forest of 1000 trees grown from samples with replacement of size 137 and the node size of 15.

This model provides a different survival function for each patient. The average number of leaves in the trees of the forest is 9.5. The c-index, computed using the out-of-bag sample of each tree, is 0.71. The most important variables are karno and cell type and also important are age and diag time.

trt	celltype	karno	diagtime	age	prior
0	0.043	0.147	0.008	0.012	0

Table 5.2: VIMP in a random survival forest

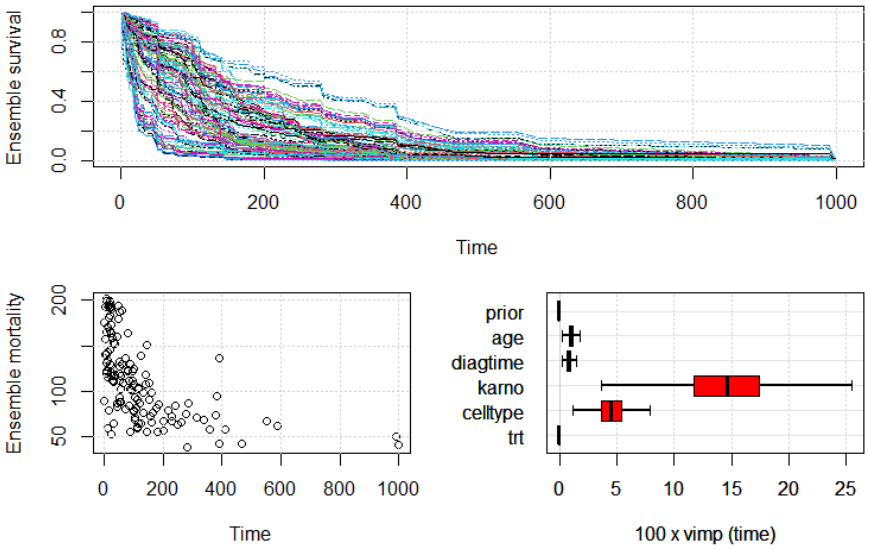


Figure 5.4: Representations of the survival function, the mortality and the VIMP computing a random survival forest

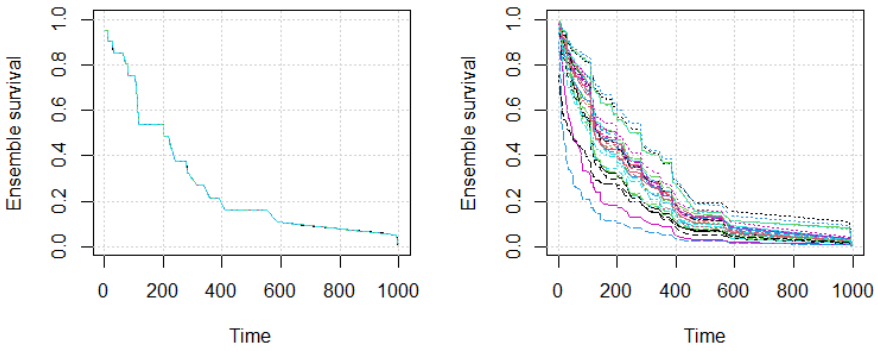


Figure 5.5: At the left the survival function for all the individuals in the same leaf of a tree (the one with minimal mortality). At the right the survival function provided by the forest for those individuals.

Example. Now we also compute an example where the real distribution of the event time is known.

We have randomly generated 25 numbers following an exponential distributions of rate $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1, in total 100 numbers generated that represent the time to event T.

```
T=rexp(100,rep(c(1/8,1/4,1/2,1),rep(25,4)))
```

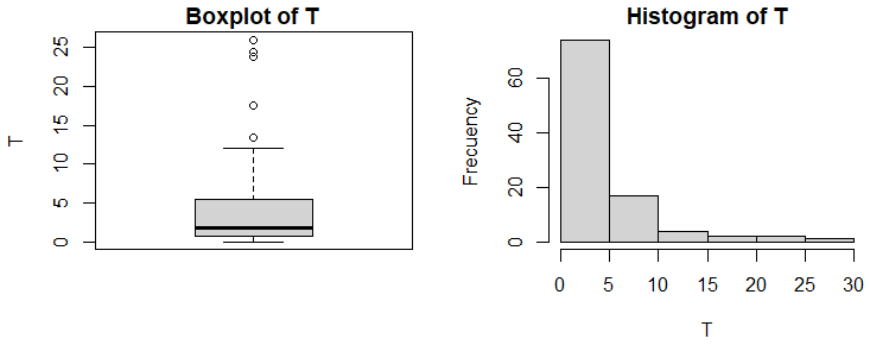


Figure 5.6: Representation of the obtained values for the time to event

We have defined a predictor variable X_1 which takes the values 1, 2, 3 and 4 for the data generated following the exponential distributions $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1 respectively.

```
X1=rep(c(1,2,3,4),rep(25,4))
```

We have defined tree noise variables generating 300 numbers from a standard normal distribution.

In order to produce censoring, we have generated 100 numbers following a uniform distribution between 0 and 15. This defines the variable L, the limits of observation.

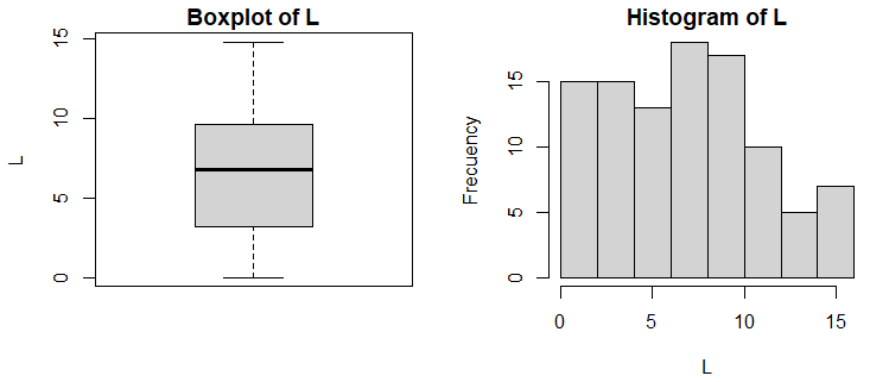


Figure 5.7: Representation of the obtained values for the limit of observation

We have defined the observed time Y as the minimum between the time to event T and the limit of observation L and δ as 1 if $Y = T$ and 0 otherwise.

Y	delta	X1	X2	X3	X4
Min. : 0.000562	Min. :0.00	Min. :1.00	Min. :-2.52981	Min. :-2.35695	Min. :-2.76453
1st Qu.: 0.551958	1st Qu.:0.75	1st Qu.:1.75	1st Qu.:-0.69035	1st Qu.:-0.79359	1st Qu.:-0.61716
Median : 1.179737	Median :1.00	Median :2.50	Median : 0.11228	Median :-0.11767	Median : 0.12063
Mean : 2.588228	Mean :0.75	Mean :2.50	Mean : 0.06484	Mean :-0.08109	Mean : 0.09673
3rd Qu.: 4.084944	3rd Qu.:1.00	3rd Qu.:3.25	3rd Qu.: 0.59910	3rd Qu.: 0.57989	3rd Qu.: 0.81908
Max. :11.975503	Max. :1.00	Max. :4.00	Max. : 3.58404	Max. : 2.30145	Max. : 2.50827

Figure 5.8: Summary of the obtained data

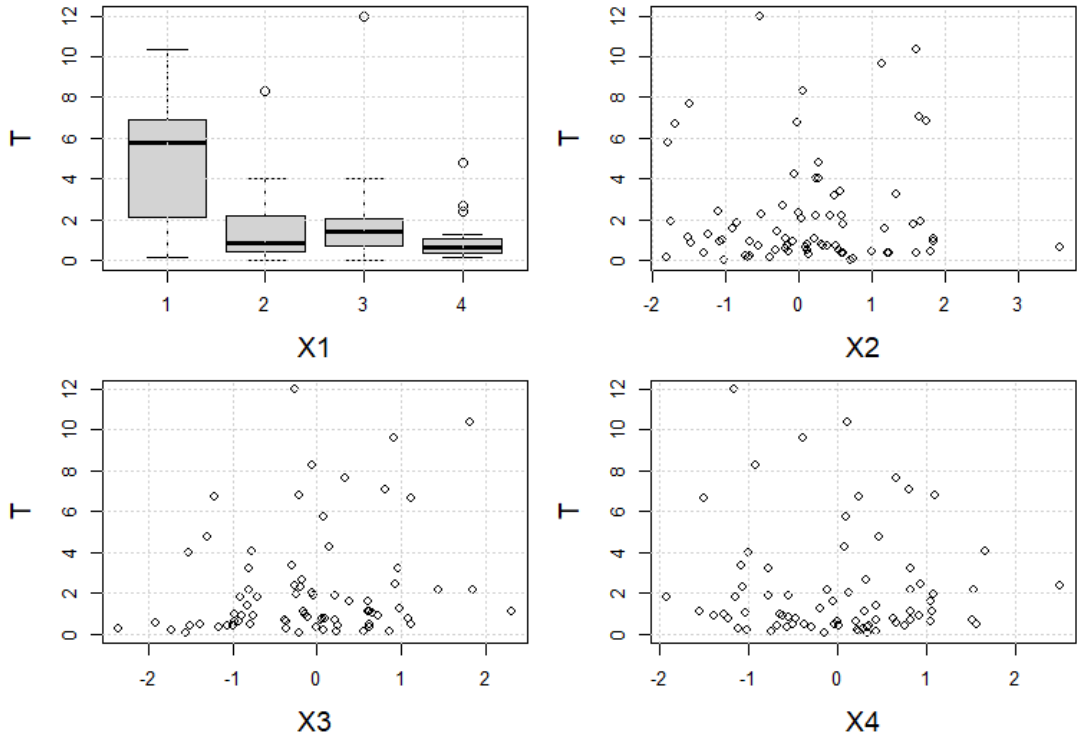


Figure 5.9: Representation of the time for non-censored data versus the predictor variables

We have computed the random survival forest algorithm using the data set obtained as described before and obtained a forest with an average number of leaves 6.717 and c-index 0.65. The variable that had the highest importance is the first one as expected and the survival function is recovered quite well for the different individuals.

X_1	X_2	X_3	X_4
0.129	-0.003	0.015	0.007

Table 5.3: VIMP in the random survival forest

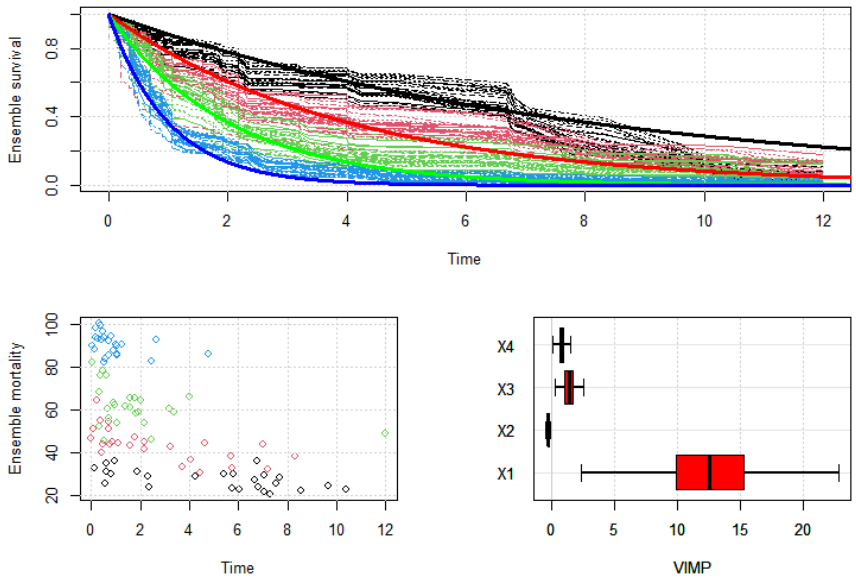


Figure 5.10: Representations of the survival function, the mortality and the VIMP computing the random survival forest. In black the individuals with $X_1 = 1$, in red the individuals with $X_1 = 2$, in green the individuals with $X_1 = 3$ and in blue the individuals with $X_1 = 4$. With the survival function graphic they are also represented the survival functions associated to the exponential distributions $\frac{1}{8}$ in black, $\frac{1}{4}$ in red, $\frac{1}{2}$ in green and 1 in blue.

Bibliography

[1] Odd Aalen. Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 3(1):15–27, 1976.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[4] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.

[5] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[6] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[7] David A Karnofsky, Walter H Abelmann, Lloyd F Craver, and Joseph H Burchenal. The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma. *Cancer*, 1(4):634–656, 1948.

[8] Michael LeBlanc and John Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

[9] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[10] R. L. Prentice. Exponential survivals with censoring and explanatory variables. *Biometrika*, 60(2):279–288, 1973.

[11] Mark Robert Segal. Regression trees for censored data. *Biometrics*, 44(1):35–47, 1988.

[12] Lev V Utkin, Andrei V Konstantinov, Viacheslav S Chukanov, Mikhail V Kots, Mikhail A Ryabinin, and Anna A Meldo. A weighted random survival forest. *Knowledge-Based Systems*, 177:136–144, 2019.

[13] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.