Statistics Surveys
Vol. 5 (2011) 44–71
ISSN: 1935-7516
DOI: 10.1214/09-SS047

A review of survival trees*

Imad Bou-Hamad

Department of Business Information and Decision Systems,
Olayan School of Business, American University of Beirut
1107, 2020 Beirut, Lebanon

Denis Larocque† and Hatem Ben-Ameur

Department of Management Sciences, HEC Montréal
3000, chemin de la Côte–Sainte–Catherine,
Montréal, Quebec, Canada H3T 2A7
e-mail: denis.larocque@hec.ca

Abstract: This paper presents a non–technical account of the developments in tree–based methods for the analysis of survival data with censoring. This review describes the initial developments, which mainly extended the existing basic tree methodologies to censored data as well as to more recent work. We also cover more complex models, more specialized methods, and more specific problems such as multivariate data, the use of time–varying covariates, discrete–scale survival data, and ensemble methods applied to survival trees. A data example is used to illustrate some methods that are implemented in R.

Keywords and phrases: Survival trees, CART, time–varying covariate, right–censored data, discrete–time, ensemble methods, time–varying effect, bagging, survival forest.

Received June 2009.

Contents

*This paper was accepted by Dorota Dabrowska, Associate Editor for the IMS.
†Corresponding author.

1. Introduction

Studies involving time–to–event data are numerous and arise in all areas of research. The Cox proportional hazard regression model and its extensions are very often used to study survival variables with censoring. These parametric (and semi–parametric) models are quite useful, for they allow simple interpretations of the covariate effects and can readily be used for inference (hypothesis testing and so on). However, such models force a specific link between the covariates and the response. Even though interactions between covariates can be incorporated, they must be specified by the analyst. Moreover, in practice, inference is often made after many models have been tried and the statistical properties of such inference after model selection are still largely unknown. When the analyst does not wish to impose a link function right from the start, more flexible approaches are available. Survival trees and forests are popular nonparametric alternatives to (semi) parametric models. They offer great flexibility and can automatically detect certain types of interactions without the need to specify them beforehand. Moreover, a single tree can naturally group subjects according to their survival behavior based on their covariates. Prognostic groups can therefore be derived easily from survival trees. Moreover, survival trees are ideal candidates for combination by means of an ensemble method and can thus be transformed into very powerful predictive tools, such as survival forests.

The development of survival trees grew from the mid–1980s up to the mid–1990s, where the goal was mainly to extend existing tree methods to the case of survival data with censoring. A review of survival trees up to 1995 appears in Leblanc and Crowley (1995). Once the basic survival tree methods were established, research moved in many different directions. One direction was to treat more complex situations such as those involving multivariate and correlated survival data (Su and Fan, 2004; Gao, Manatunga and Chen, 2004; Fan et al., 2006; Gao, Manatunga and Chen, 2006; Fan, Nunn and Su, 2009). Another direction was to study the use of ensemble methods with survival trees (Ishwaran et al., 2004; Hothorn et al., 2004, 2006; Krętowska, 2006, 2010; Ishwaran et al., 2008). Yet another dealt with specific topics related to survival studies, such as time–varying covariates and time–to–event variables measured on a discrete scale (Bacchetti and Segal, 1995; Huang, Chen and Soong, 1998;

Xu and Adak, 2001, 2002; Yin and Anderson, 2002; Bou-Hamad et al., 2009; Bou-Hamad, Larocque and Ben-Ameur, 2011).

The rest of this section describes the basic tree methodology and the survival data setup. Section 2 focuses on the basic survival–tree methodologies. Ensemble of survival trees are discussed in Section 3. In Section 4, more recent extensions of survival trees are presented. A data example illustrates some of the methods in Section 5. Finally, some concluding remarks along with possibilities for future research are given in Section 6. A more complete discussion about pruning and the selection of a final tree is deferred to the Appendix.

1.1. Basic tree building method

Initially, tree–based methods were developed to model a categorical or a continuous outcome using a set of covariates from a sample of complete data. They were introduced by Morgan and Sonquist (1963) but really became popular in the 1980s due in great part to the development of the CART (Classification and Regression Tree) paradigm described in the monograph by Breiman et al. (1984). Assuming that the reader is familiar with the basic ideas and terminology of tree–based methods, only a brief description is provided here. The basic idea of a tree is to partition the covariate space recursively to form groups (nodes in the tree) of subjects which are similar according to the outcome of interest. This is often achieved by minimizing a measure of node impurity. For a categorical response, the Gini and the entropy measures of impurity are popular, while the sum of squared deviations from the mean is most often used with a continuous outcome.

The basic approach focuses on binary splits using a single covariate. For a continuous or an ordinal covariate X, a potential split has the form $X \leq c$ where c is a constant. For a categorical covariate X, a potential split has the form $X \in \{c_1, \ldots, c_k\}$ where $c_1, \ldots, c_k$ are possible values of X. The typical algorithm starts at the root node with all observations; performs an exhaustive search through all potential binary splits with the covariates; and selects the best one according to a splitting criterion such as an impurity measure. In the CART approach, the process is repeated recursively on the children nodes until a stopping criterion is met (often until a minimum node size is attained). This produces a large tree that usually overfits the data. A pruning and selection method is then applied to find an appropriate subtree. Alternatively, an ensemble of trees can be used which avoids the problem of selecting a single tree of appropriate size. Appropriate node summaries are usually computed at the terminal nodes to interpret the tree or obtain predicted values. The node average is typically used for a continuous outcome, whereas, for a categorical outcome, the node proportions of each value are reported. The most frequent value at a node can be used if a single prediction is needed. For a survival outcome, the Kaplan–Meier estimate of the survival function in the node can be reported.

1.2. Survival data description

We begin by describing the basic setup which leads to the development of survival trees. We denote by U the true survival time and by C the true censoring time. The observed data is then composed of $\tau = \min(U, C)$, the time until either the event occurs or the subject is censored; $\delta = I(U \leq C)$, an indicator that takes a value of 1 if the true time–to–event is observed and 0 if the subject is censored; and $\mathbf{X} = (X_1, \ldots, X_p)$, a vector of p covariates. Data is available for N independent subjects $(\tau_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, N$. The basic setup assumes that the covariate values are available at time 0 for each subject. Thus, only the baseline values of a time–varying covariate are typically used. The inclusion of the multiple values of time–varying covariates will be discussed in Section 4.2. Multivariate and correlated survival data will be the topic of Section 4.1.

2. Survival tree building methods

The early idea of using tree–structured data analysis for censored data can be traced back to Ciampi et al. (1981) and Marubini, Morabito and Valsecchi (1983). However, the first paper containing all the elements of what would become survival trees was written by Gordon and Olshen (1985).

In this section, we present the splitting criteria proposed over the years and discuss briefly the choice of a final tree. We also present some variants and related methods as well as the few studies that have compared some of the tree–building procedures.

2.1. Splitting criteria

The idea behind the splitting criterion proposed by Gordon and Olshen (1985) was to force each node to be more homogeneous as measured by a Wasserstein metric between the survival function obtained from the Kaplan–Meier estimator at the node and a survival function that has mass on at most one finite point. Although this particular splitting criterion did not gain much popularity, it laid ground for the work that followed. Indeed, Gordon and Olshen (1985) mention the possibility of using the logrank statistic or a parametric likelihood ratio statistic to measure the "distance" between the two children nodes and these ideas have been used widely in subsequent work.

Ciampi et al. (1986) then suggested using the logrank statistic to compare the two groups formed by the children nodes. The retained split is the one with the largest significant test statistic value. The use of the logrank test leads to a split which assures the best separation of the median survival times in the two children nodes. Ciampi et al. (1987) subsequently proposed a general formulation based on using the likelihood ratio statistic (LRS) under an assumed model to measure the dissimilarity between the two children nodes. As for the logrank statistic above, it is clear that the larger the statistic, the more dissimilar the two nodes. They discuss more specifically two possibilities: an exponential model

and a Cox proportional hazard model. Hence, this approach relies on the assumptions related to the chosen model. For instance, with the Cox model, the proportional hazard assumption implies that the hazard function in the right node is proportional to the one in the left node. Davis and Anderson (1989) use a splitting criterion based on an exponential model log–likelihood which is equivalent to the LRS dissimilarity measure under the exponential model. Continuing in the same direction, Ciampi et al. (1988) and Ciampi, Thiffault and Sagman (1989) mention the possibility of using the logrank and Wilcoxon–Gehan statistics as dissimilarity measures and the Kolmogorov–Smirnov statistic to compare the survival curves of the two nodes. Segal (1988) also adopts a between–node separation (dissimilarity measure) approach based on the Tarone–Ware class of two–sample statistics for censored data. With appropriate choices of weights, this class encompasses many well–known test statistics, such as the logrank and Wilcoxon–Gehan statistics. Leblanc and Crowley (1993) also use the logrank statistic as a splitting criterion, but they introduce a new method for pruning and selecting a final tree built around a measure of split–complexity (see Appendix).

In their discussion, Therneau, Grambsch and Fleming (1990) mention that martingale residuals from a null Cox model could be used as the outcome for a regression tree algorithm. The advantage of this approach is that existing regression tree software could be used directly with the modified outcome. Keles and Segal (2002) provide an analytic relationship between the logrank and martingale residuals sum–of–squares split functions. However, their approach is based on the idea that the residuals are recomputed at each node, thus ruling out the direct use of regression tree software. They show that the two splitting criteria are approximately equivalent when the survival time is independent of the covariate, but not in the general case. Loh (1991) and Ahn and Loh (1994) propose two splitting criteria based on residuals obtained from fitting a Cox model with one covariate at a time. The basic idea consists in studying the patterns of the Cox model residuals along each covariate axis and then selecting the splitting covariate whose axis patterns appear the least random. The degree of randomness of the residuals is quantified by dividing the observations in the parent node into two classes along each covariate and is measured by the two–sample t–test.

By exploiting an equivalence between the proportional hazard, full likelihood model and a Poisson likelihood model, Leblanc and Crowley (1992) come up with a splitting criterion based on a node deviance measure between a saturated model log–likelihood and a maximized log–likelihood. With this method, the unknown full likelihood is approximated by replacing the baseline cumulative hazard function by the Nelson–Aalen estimator. The advantage of this method is that it can be implemented easily in any recursive partitioning software for Poisson trees such as the rpart package in R (Therneau and Atkinson, 2010).

Zhang (1995) proposes an impurity criterion which combines two separate impurity measures, one for the observed times and one for the proportion of censored observations. Molinaro, Dudoit and van der Laan (2004) propose a unified strategy for building trees with censored data. Their approach is based on defin-

ing an observed data–world (with censoring) loss function by weighting a full data–world (without censoring) loss function. Each non–censored observation is weighted by the inverse probability of censoring (IPC) given the covariates. Since the usual regression tree methods use the node variance as the impurity measure, Jin et al. (2004) propose a splitting rule based on the variance of survival times. But since mean and variance survival times are affected by the censored observations, they propose using a restricted time limit to compute the variance. Finally, Cho and Hong (2008) propose using the L_1 loss function to build a median survival tree. To compute the loss function, the censored observations are replaced by their expected values, conditional on the fact that the time is greater than the censored time.

2.2. Selection of a single tree

One important aspect when building a single tree is deciding when to stop splitting and hence select a specific tree as the final model. If too large, trees will tend to overfit the data and thus fail to generalize well to the population of interest. If too small, they might miss important characteristics of the relationship between the covariates and the outcome. There are basically two approaches to the selection of a final tree. The first one is a backward method which builds a large tree and then selects an appropriate subtree by pruning some of its branches. The second one is a forward method which uses a built–in stopping rule to decide when to stop splitting a node further. However, the use of a single tree has been largely replaced by ensemble of trees which often produce more powerful predictive models and which also avoid the problem of selecting a single tree. But a single tree can still be of interest to gain insight about the data since it can be easily interpreted. Consequently, the discussion about the selecting a single tree is deferred to the Appendix.

2.3. Some variants and related methods

The RECPAM (Recursive Partition and Amalgamation) method introduced in Ciampi et al. (1988) allows a new feature to appear in the classical tree; see Ciampi, Negassa and Lou (1995) for a complete description. Their method shares the basic characteristics of regular trees in the sense that it builds a large tree, prunes it, and then selects one member in the sequence as the final tree. However, it allows a further step, the amalgamation step, where similar terminal nodes are grouped together. The amalgamation algorithm proceeds like a pruning and selection algorithm inasmuch as it recursively amalgamates the two terminal nodes which are the most similar to create a sequence of nested partitions from which one final partition will be selected. In the end, the partition of the covariate space may not necessarily be that of a tree, since widely separated terminal nodes may end up being grouped together. But it may bring the number of groups down to a more easily interpretable size. In their data example, Fan et al. (2006) use an amalgamation algorithm to bring

the 12 terminal nodes of their final tree down to five interpretable prognosis groups.

A similar idea of building a tree and then grouping together terminal nodes which are similar with respect to the survival profiles is proposed in Tsai et al. (2007). The grouping of the terminal nodes of the final tree is achieved with an agglomerative hierarchical clustering method. The method developed by Leblanc and Crowley (1995) also breaks away from the tree structure and can build proportional hazard models with piecewise constant relative risk functions. Adapting the ideas of Logical Analysis of Data or LAD (Hammer and Bonates, 2006; Kronek and Reddy, 2008) propose the LASD (Logical Analysis of Survival Data) method that automatically detects good patterns of covariates to predict the survival function. Su and Tsai (2005) propose a hybrid approach using a tree structure to expand the Cox proportional hazard model. Finally, Kręłowska (2004) uses the concept of dipolar trees to build survival trees. At a given node, the basic idea is to compare all pairs of observations and decide, based on a criterion, which pairs should be together and which should be separated after the split. Then the linear combination of the covariates (a hyperplane) that preserves this "ideal" split is found according to a dipolar criterion function. As such, the splits in the trees are formed by linear combinations of the covariates.

2.4. Comparison of methods

A large scale simulation study comparing the numerous pruning and selection methods has yet to appear but some limited empirical work is available. To investigate the performance of some tree–size selection methods within the RECPAM framework, Negassa et al. (2000, 2005) have studied the performance of four model selection approaches: cross–validation, cross–validation with the 1 SE rule (Breiman et al., 1984), automatic elbow rule, and minimum AIC. They conclude that none of these approaches exhibits a uniformly superior performance over all scenarios. They also propose a two–stage method where cross–validation is used in the first stage and the elbow approach in the second. This method performed well in their simulation.

A large scale comparison of the numerous splitting criteria has also yet to appear. Some limited results do appear in Radespiel-Tröger et al. (2003) and Radespiel-Tröger et al. (2006). The first paper uses a real data set as well as a single tree–structured data generating process with five terminal nodes and sample sizes of 250 but with many variations of censoring distributions and terminal node hazards. Having compared many splitting methods, the authors conclude that adjusted and unadjusted logrank statistic splitting with pruning, exponential loss splitting with pruning, and adjusted logrank statistic splitting without pruning perform best. Radespiel-Tröger et al. (2003) use bootstrap samples from a real data set to perform their simulation study. Their results show that adjusted logrank statistic splitting without pruning gives the best results.

3. Ensemble methods with survival trees

Trees are known for their instability, in the sense that small perturbations in the learning sample can induce a large change in their predictive function. Bagging and random forests, proposed by Breiman (1996, 2001), are simple but ingenious solutions to this problem that basically reduce the variance of a single tree and enlarge the class of models. In fact, bagging is one particular case of random forests. The basic algorithm works by drawing B bootstrap samples from the original data and growing a tree for each of them without pruning. A final prediction is then obtained by averaging the predictions from each individual tree. The general random forest algorithm grows each tree by selecting a random subset of covariates at each node. Bagging is then just the special case where all covariates are retained at each node. Siroky (2009) and Verikas, Gelzinis and Bacauskiene (2011) provide recent discussions on random forests. Moreover, random forests are part of the family of ensemble methods for which a survey appears in Rokach (2008).

Dannegger (2000) and Benner (2002) describe applications of bagging with survival trees. Ishwaran et al. (2004) propose a forest of relative risk trees using the tree–building method introduced in Leblanc and Crowley (1992), a model which assumes proportional hazards. For any given covariate $\mathbf{x}$, each tree (for $b = 1, \ldots, B$) produces a relative risk value $R^{(b)}(\mathbf{x})$ compared to the mean unit in the study. They define the ensemble relative risk for $\mathbf{x}$ to be $R_e(\mathbf{x}) = 1/B \sum_{b=1}^{B} R^{(b)}(\mathbf{x})$.

Hothorn et al. (2004) propose a general bagging method for an arbitrary tree growing algorithm but use the Leblanc and Crowley (1992) method for their practical implementation. However, their method differs in the way they aggregate the individual trees. To obtain an estimate of the survival function at a covariate $\mathbf{x}$, they form a new set of observations by collecting from each tree and from the bootstrap sample used to build the tree all the observations that fall into the same terminal node as $\mathbf{x}$. They then compute the Kaplan–Meier estimate using this set of observations. Thus, they end up with a conditional survival function which is more informative than a single prediction like a median survival time or a relative risk compared to a mean unit. Their method is implemented in the R package ipred (Peters and Hothorn, 2009).

Hothorn et al. (2006) propose a random forest method to build a survival ensemble for the log–survival time. Their approach is based on the general Molinaro, Dudoit and van der Laan (2004) framework. The estimated inverse probability of censoring (IPC) weights are used as sampling weights to draw each bootstrap sample and a tree is built for each of them. With the quadratic loss, a prediction of the mean log–survival time at a covariate $\mathbf{x}$ is given by the average survival time of the terminal node corresponding to $\mathbf{x}$. The ensemble prediction of the mean log–survival time is then obtained as a weighted average, over all trees, of these predictions. Their method is implemented in the R package party. They also investigate a gradient boosting algorithm where a tree can act as the base learner, but they look instead at the use of component–wise least squares. Hence, this particular boosting method is not really an extension

of survival trees. Along the same lines, Ridgeway (1999) and Benner (2002) also propose a boosting algorithm with different base learners.

Ishwaran et al. (2008) introduce a general random forest method, called random survival forest (RSF), coupled with a new algorithm for imputing missing values. They investigate four different criteria based on versions of the logrank–statistics and conservation–of–events principle. To obtain a prediction at a given $\mathbf{x}$, the Nelson–Aalen estimates of the cumulative hazard function at each node are averaged. Uniform consistency of RSF, under the assumption of a discrete feature space, is established in Ishwaran and Kogalur (2010a). Since a forest is mainly a black box, quantifying a covariate importance in a forest is a difficult but important problem. In the original random forests paper, Breiman (2001) propose a variable importance measure (VIMP) which works by examining the prediction error increase when noise is added to a covariate. Ishwaran et al. (2010) propose a variable selection method for survival data and studied it through RSF. This method is based on the concept of minimal depth of a maximal subtree which basically assesses the importance of a covariate by measuring how deep in a tree the first split based on it occurs. The idea being that covariates splitting higher in a tree are more important. All these methods are implemented in the R package randomSurvivalForest (Ishwaran and Kogalur, 2010b).

Kręłowska (2006, 2010) studied forests of dipolar survival trees. As in Hothorn et al. (2004), the final Kaplan–Meier estimate is computed by using the set of aggregated observations from all individual trees.

Eckel et al. (2008) compare proportionnal hazard models, survival forests, and a bundling method with a data set of melanoma patients. The bundling method combines the Cox model with a tree including the linear predictor of a Cox model as an additional predictor, thus expanding the candidate splits. The final predictions are obtained from aggregated trees. They conclude that the three methods are on par for that data set.

4. Extensions of the basic methods

The last sections presented the developments in survival trees and related methods for the basic setup involving a univariate survival outcome with independent data and without time–varying covariates. Extensions to more complex situations began to appear in the mid–1990s. This section will present these developments in a thematic fashion. Extensions to multivariate and correlated data will be presented first, followed by specialized topics such as time–varying covariates, time–to–event variables measured on a discrete scale, other forms of censoring and the treatment of missing values.

4.1. Multivariate and correlated data

A natural extension of univariate survival–tree methods is to consider multivariate or correlated survival outcomes. Suppose that there are N clusters in

the data. Using the same notation as in Section 1.2, the available data are $(\tau_{ij}, \delta_{ij}, \mathbf{X}_{ij})$, where the (ij) subscript indicates the observations for the unit j in cluster i, $j = 1, \ldots n_i$, $i = 1, \ldots, N$. Independence is assumed across clusters but the observations within a cluster are possibly correlated. The goal is to build a survival tree by taking into account the intra–cluster correlation. The marginal and random effect (frailty) models are the two main approaches taken in handling correlated survival outcomes and both have been adapted to the construction of survival trees.

Su and Fan (2004); Gao, Manatunga and Chen (2004); Fan, Nunn and Su (2009) use the frailty approach where the intra–cluster dependence is modeled by a multiplicative random–effect term. More specifically, the following formulation of the hazard function is the starting point of their method:

$$h_{ij}(t|\mathbf{X}_{ij}, w_i) = h_0(t) \exp(\mathbf{X}_{ij}\boldsymbol{\beta}) w_i$$

where h_0 is an unspecified baseline hazard function and w_i is a frailty term for cluster i that follows some specified distribution. The gamma distribution is assumed in these papers. To define a splitting criterion, Su and Fan (2004) use the deviation of an integrated log–likelihood, Fan, Nunn and Su (2009) use a score test on the splitting variable derived from the integrated log–likelihood and Gao, Manatunga and Chen (2004) use a standardized estimate of the splitting variable parameter obtained from a profile log–likelihood.

Fan et al. (2006) use the marginal approach where the dependence structure is left unspecified. More precisely, they consider the following Cox model to evaluate a single binary splitting variable $C \in \{0, 1\}$ defined through one of the covariate:

$$h_{ij}(t|C_{ij}) = h_0(t) \exp(\beta C_{ij})$$

where h_0 is an unspecified baseline hazard function. By using a consistent estimate of the variance structure of the score function, they obtain a score test of the null hypothesis $H_0 : \beta = 0$ which acts as their splitting criterion. This test is a robust two–sample logrank statistic and their whole methodology is a generalization of the Leblanc and Crowley (1993) method. One advantage of this approach over the frailty approach is that it does not require iterative procedures, since the robust logrank statistic has a closed–form expression. Gao, Manatunga and Chen (2006) use a similar approach based on the marginal distribution of survival time but assume a global proportional hazard structure for the whole tree. Contrary to the usual approaches, the whole data set is used at each split.

4.2. Specific topics: Time–varying effects and covariates, discrete–time survival outcome and other types of censoring

Almost all survival tree methods have been developed under the basic setup described in section 1.2 and so they include neither time–varying effects nor

time–varying covariates. Moreover, no method specifically adapted to discrete–time survival data has been proposed until very recently (Bou-Hamad et al., 2009).

Given that time–varying covariates are common in practice, only the difficulties associated with their use can explain the sparsity of literature on tree–based methods which treat this topic. In the context of regression trees for longitudinal data, Segal (1992) discusses issues about time–varying covariates and points out that no convincing technique for defining splits on them has been developed. One possibility is to replace each time–varying covariate by estimated parameters that summarize its relation to time. For instance, if the values of a time–varying covariate of an individual are regressed over time, then the slope and intercept could be used in the tree–growing process instead of the original values. But this is not really satisfactory for two reasons. First, there is no guarantee that the covariate is linearly related to time. Second, the number of repeated measures on an individual is generally too small to allow precise regression estimates.

The first studies dealing with time–varying covariates with survival trees are the ones by Bacchetti and Segal (1995) and Huang, Chen and Soong (1998) The solution proposed by Bacchetti and Segal (1995) is to allow the decomposition of each subject into pseudo–subjects defined by the tree's splitting rules. Assume that $x(t)$ is a time–varying covariate. If the splitting rule at a node is $x(t) \leq c$, then the time window where this condition is true would go to one node and the time window where it is false would go to the other node. Hence, a subject could be split into two pseudo–subjects that could be split further at lower nodes. In the end, a subject could end up in many different terminal nodes. However, at any given time, each subject can be classified into one and only one terminal node. In order to achieve this, Bacchetti and Segal (1995) use modified two–sample test statistics that can accommodate left–truncated data. Huang, Chen and Soong (1998) use a similar approach in which subjects can be split across many nodes as a function of time but with a more structured model. Their splitting criterion is built around the log–likelihood of a model which assumes that the distribution of the survival time for a subject is given by a piecewise exponential distribution.

Xu and Adak (2001, 2002) use only time independent covariates, but these are allowed to have time–varying effects. With their methods, the tree is used only to find splits for the time variable in order to locate those time values where effect changes occur. The resulting tree partitions the time into time intervals and a Cox proportional hazard model is used to model the covariates. Hence, this model fits an adaptive piecewise Cox model by letting the tree algorithm find the intervals.

Bou-Hamad et al. (2009) propose a method specifically adapted to discrete–time survival outcomes. Their splitting criterion is based on the log–likelihood of a very flexible discrete–time model which reduces to the entropy criterion for a categorical response when no censored observations are present. Moreover, this method directly allows time–varying effects for the covariates. Bou-Hamad, Larocque and Ben-Ameur (2011) generalize this approach so as to incorporate

time–varying covariates. This is achieved by allowing subjects to be split across different nodes depending on the time period, as in Bacchetti and Segal (1995). Hence, this method allows, simultaneously, both time–varying effects and time–varying covariates. These two papers also investigate the use of bagging and random forests to produce aggregate estimations of the discrete conditional risk and survival functions.

By allowing left–truncation, Bacchetti and Segal (1995) is a rare example of a method that goes beyond right–censoring. Another one is Yin and Anderson (2002) which extends the exponential tree method of Davis and Anderson (1989) to interval–censored data.

4.3. Missing values

Missing data is an important problem in practice. Several approaches have been investigated in the context of classification or regression trees. These include the use of surrogate splits, using imputation methods and treating missing values as a "new" value. In principle, these methods can also be used with survival trees. A recent point of entry in the literature about missing values and trees is Ding and Simonoff (2010). However, there are very few specific investigations on the treatment of missing values with survival trees.

Ishwaran et al. (2008) argue that the traditional surrogate split approach is possibly not well–adapted to the forest paradigm because finding a surrogate split is computationally intensive and because a forest typically selects a small subset of covariates at each node of a tree to find the best split. Hence, finding a good surrogate split with only a few candidate covariates may be problematic. This is why they introduced a new missing data algorithm built in their RSF method. The basic idea is to impute the missing values at the parent node prior to splitting using only the in–bag data. Thus, the out–of–bag (OOB) data are left untouched and the prediction error estimates based on them are not optimistically biased. Once, the missing values are imputed, the splitting proceeds as usual. They also propose to iterate the whole process. Their method is implemented in the R package randomSurvivalForest.

For a single survival tree, Wallace, Anderson and Mazumdar (2010) propose a multiple imputation technique. The basic idea is to impute the missing values of a covariate by building a tree using the other covariates as predictors. By adding noise, multiple imputed values can be obtained. To select the best split at a given node, the splitting statistic is computed as usual for covariates without missing values. For a covariate which required imputation, the splitting statistic is computed for each set of imputed values. The median of these splitting statistics is then used as the final splitting statistic for this covariate. Tree–building can then proceed as usual. They performed a simulation study where they compared different versions of their method to a benchmark strategy that uses only the complete data to grow the tree and to the use of surrogate splits. The results suggest that their proposed method is better than the others for identifying the correct tree structure while remaining competitive with respect to prediction accuracy.

5. A data example

In this section, we illustrate some aspects of survival trees and forests with the well–known PBC data set. A description an analysis of the data set is presented in the monograph by Fleming and Harrington (1991). The data are from the Mayo Clinic randomized placebo controlled trial of the drug D–penicillamine, for the treatment of primary biliary cirrhosis (PBC) of the liver, conducted between 1974 and 1984. The 312 patients that participated in the randomized trial are used in this example. Subject survival (in days) since registration in the trial is the outcome of interest and censoring is due to either liver transplantation or study analysis time (in July 1986). Seventeen covariates are available but only the twelve "inexpensive, non–invasive and readily available" covariates are retained for this example as it was done in the analysis presented in Section 4.4 of Fleming and Harrington (1991). These are:

1. Drug: 1=D–penicillamine, 0=placebo.
2. Age: age in years.
3. Sex: 0=male, 1-female.
4. Ascites: presence of ascites (0=no, 1=yes).
5. Hepatom: presence of hepatomegaly (0=no, 1=yes).
6. Spiders: presence of spiders (0=no, 1=yes).
7. Edema: presence of edema (0=no edema and no diuretic therapy for edema; 0.5=edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1=edema despite diuretic therapy.
8. Bili: Serum bilirubin, in mg/dl.
9. Albumin: in gm/dl.
10. Alkphos: alkaline phosphatase, in U/liter.
11. Platelet: platelet count, in number of platelets per–cubic–milliliter of blood divided by 1000.
12. Protime: prothrombin time, in seconds.

As in Fleming and Harrington (1991), we replace the four missing platelet counts by the 257, the median of the other observations, and, as such, we are using the same data set as them. All computations are performed with R (R Development Core Team, 2010).

5.1. A single tree

First, a single tree is built using the default settings in the rpart package. With a survival outcome, the splitting criterion used by rpart is equivalent to the one of Leblanc and Crowley (1992). The final pruned tree selected by cross–validation has 12 terminal nodes and is shown in Figure 1. The first split is based on the variable Bili. Subjects with a value less than 2.25 go to the left node and those with a value greater or equal to 2.25 go to the right node. Each terminal node contains the following information: a letter from A through L used to match the node with the survival curves of Figure 2, the node sample size and the

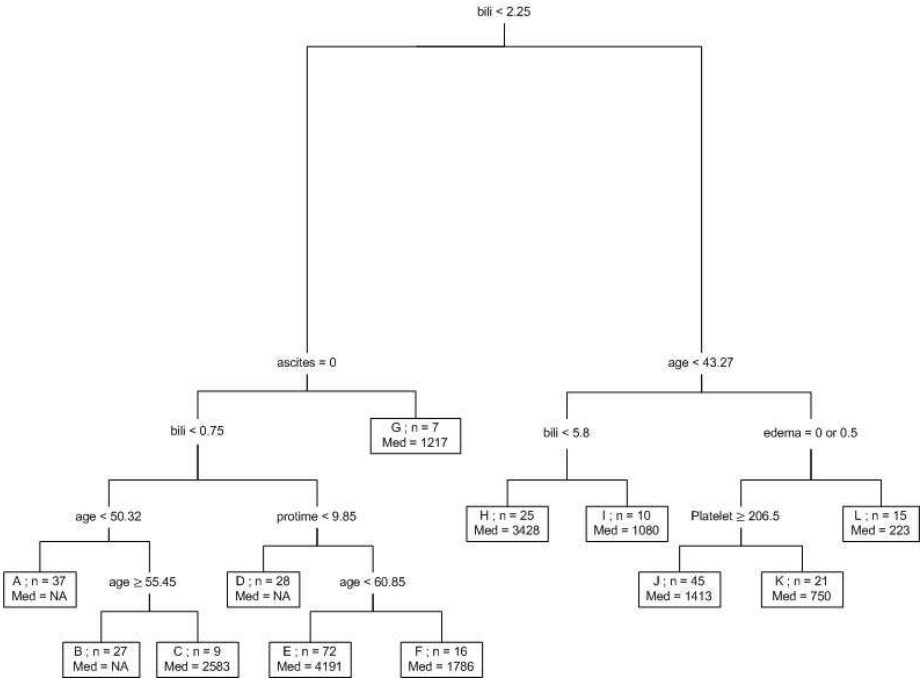

FIG 1. *Single tree for the PBC data example.*

estimated survival median in the node. For instance, Node E has 72 subjects and a median survival time of 4191 days. The Kaplan–Meier estimates of the survival function of the terminal nodes are presented in Figure 2. The letters placed at the median survival time (except for nodes A, B and D for which the median is not defined), refer to the terminal nodes. The overall Kaplan–Meier curve of the 312 subjects is under the letter "O". After the first split, we can see that the subjects with a value of Bili greater or equal to 2.25 (especially those in nodes I, J, K and L) tend to have lower survival time. On the other hand, the terminal nodes A B, D and E seem to have the most favorable survival patterns. For instance, the largest node (node E) is formed by subjects with Bili $\in [0.75, 2.25)$, Ascites=0, Protime$\geq$9.85 and Age$<$60.85. However, we will see that in this example, a single tree is not such a good predictor of the survival function compared to aggregation methods.

5.2. Bagging, forests and comparison of methods

Four other methods are used to obtain prediction of the survival curves. The first one is not using any covariate information and is simply the Kaplan–Meier estimate of the sample. The second method is a basic Cox model including the main effects of all covariates. No transformations are performed. These two methods serve as benchmarks. The first aggregation method is the bagging method of

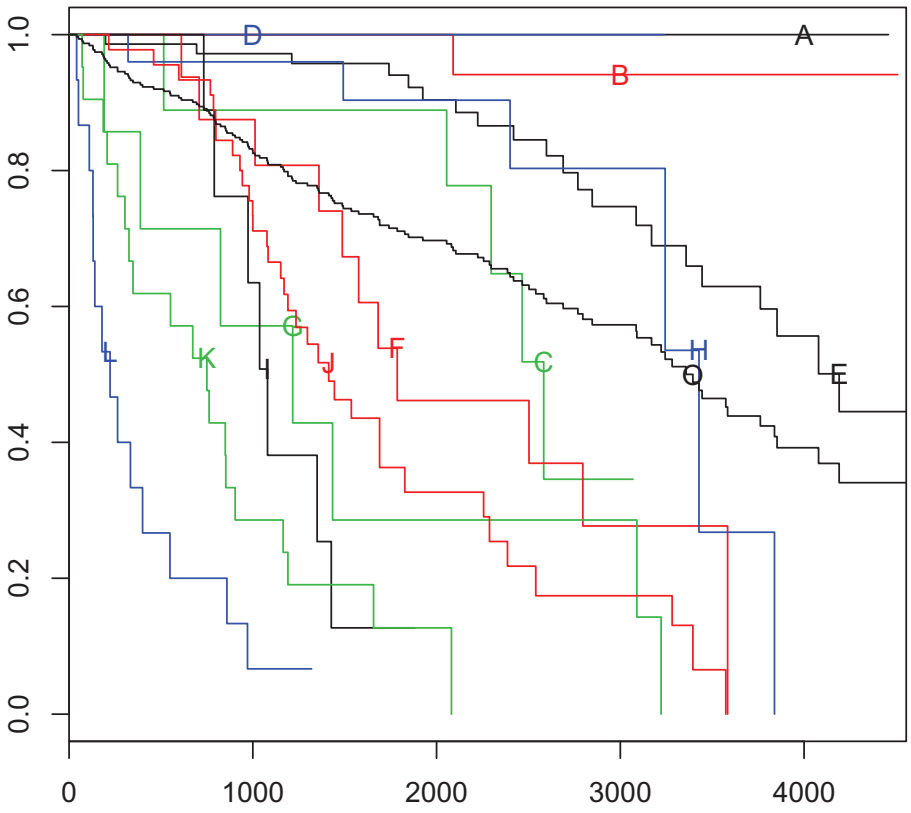

FIG 2. *Kaplan–Meier survival estimate in each node of the single tree of Figure 1 for the PBC data example. The letters A through L correspond to the terminal nodes of the tree as given in Figure 1. The line "O" is the Kaplan-Meier curve of the whole sample.*

Hothorn et al. (2004) as implemented in the package ipred. Basically, it aggregates many trees built with rpart. Here, we use 1000 trees. The second aggregation method is the random survival forest (RSF) approach of Ishwaran et al. (2008) as implemented in the package randomSurvivalForest. The logrank splitting rule is used and once again, 1000 trees are built for each forest.

Our goal is to compare the performance of the five methods to estimate the survival function of a new subject. However, there is no universal and widely accepted performance measure to assess the accuracies of estimated survival functions. The integrated Brier score (Graf et al., 1999) is a popular measure of performance and we use it in our calculations. Using the notation of Section 1.2, let $\hat{S}(t|\mathbf{X})$ denote the estimated survival function at time t of a subject with covariate vector $\mathbf{X}$. This estimate may come from any models. Let $\hat{G}(t)$ denote the Kaplan-Meier estimate of the censoring survival function. This is simply the Kaplan-Meier estimate based on $(\tau_i, 1-\delta_i)$, $i = 1, \ldots, N$. The Brier score at

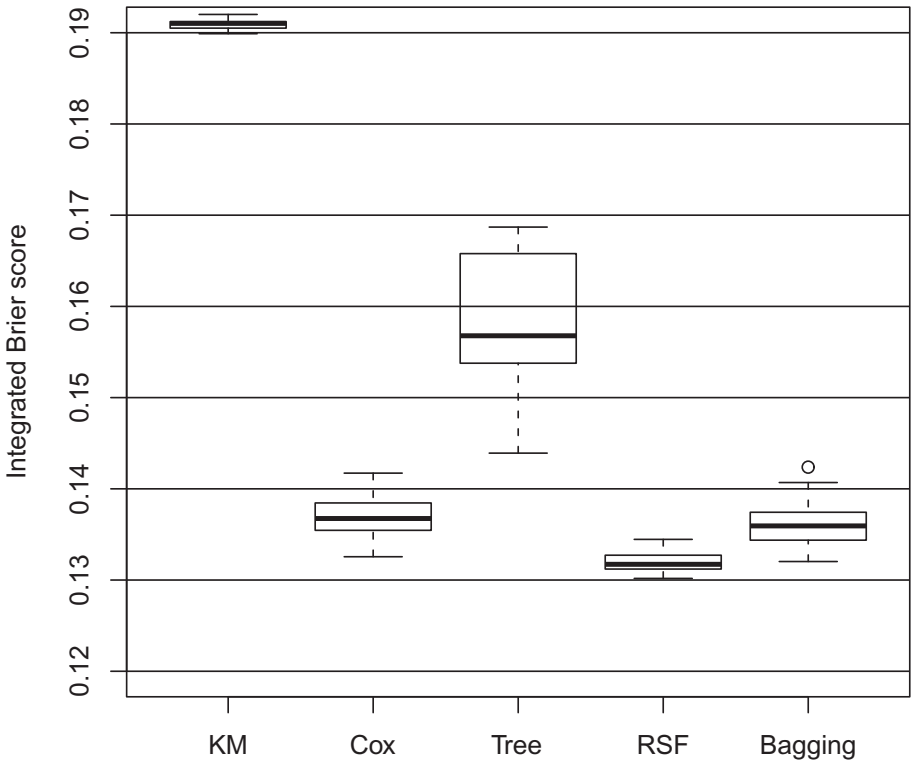

FIG 3. *Integrated Brier Score across the 20 runs of 10–fold cross–validation for the PBC data example.*

time t is given by

$$\mathrm{BS}(t) = \frac{1}{N}\sum_{i=1}^{N}\Big((\hat{S}(t|\mathbf{X_i}))^2 I(\tau_i \leq t \text{ and } \delta_i = 1)\hat{G}^{-1}(\tau_i) + (1-\hat{S}(t|\mathbf{X_i}))^2 I(\tau_i > t)\hat{G}^{-1}(t)\Big).$$

The integrated Brier score is then given by

$$\mathrm{IBS} = \frac{1}{\max(\tau_i)}\int_0^{\max(\tau_i)} \mathrm{BS}(t)dt,$$

and lower values indicate better predictive performances.

We perform 10–fold cross–validation (cv) 20 times. For instance, 200 RSF (each one with 1000 trees) are built during the process. We then obtain 20 estimates of the IBS for each methods. The box–plots of the 20 values of IBS for each methods are presented in Figure 3. We see that the best result is obtained for RSF and that bagging and the Cox model give a similar performance. It is

also clear that a single tree does not perform well in this example and that its performance varies a lot more than the other methods across the 20 cv runs. This illustrates the potential instability of single trees and the fact that aggregating many trees can solve this problem.

5.3. Variable importance and visualization of a covariate effect

We pursue the analysis by studying the importance of the covariates. Our approach is computer–intensive but our results will be compared to some readily available importance measures arising from RSF. Each covariate is removed one at a time. We then estimate, again by repeating 10–fold cv 20 times, the IBS of each method without the covariate. Thus, the whole process is repeated twelve times (one time for each covariate). The average of the 20 estimated IBS then serves as the final performance measure. Figure 4 presents the percent increase (or decrease if negative) in IBS when a single covariate is removed compared to the model with all covariates. Obviously, the Kaplan-Meier method is not there

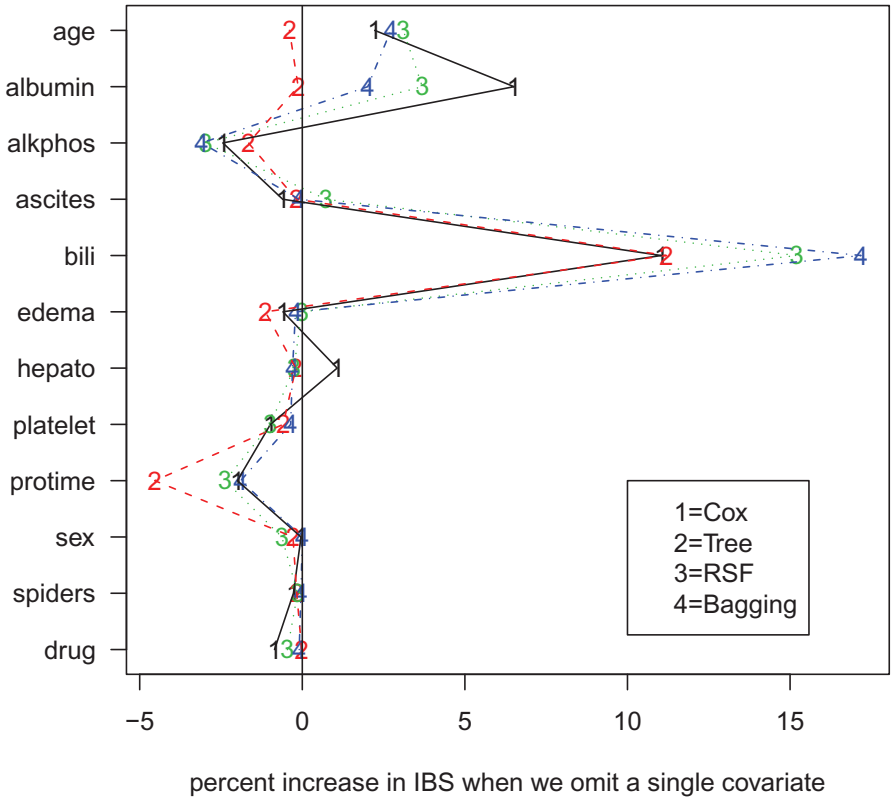

FIG 4. *Variable importance for the four models when omitting a single covariate for the PBC data example.*

since it is not using any covariates in the first place. We clearly see that, for all methods, the covariate Bili seems to be the most important. Not using this covariate increases the IBS by 17.2% and 15.2% for bagging and RSF respectively. The increase is less pronounced but still noticeable (more than 11%) for the single tree and the Cox models. Except for the single tree model, Age and Albumin seem to be the following covariates in terms of importance but their increase in IBS is less than 5% (except for Albumin with the Cox model where it reaches 6.5%). It is interesting to note that, with this measure, the treatment effect (covariate Drug) is not important for all four models.

Less computer–intensive variable importance measures are available in randomSurvivalForest. The first two depend on a performance measure (Ishwaran et al., 2008) while the third one, the minimal depth of a maximal subtree, does not (Ishwaran et al., 2010). Using only, the RSF model, we compare the omit one covariate importance measure presented above with these three. In randomSurvivalForest, Harrell's concordance index, the C–index, is used to estimate the performance (Harrell et al., 1982). It is basically an estimate of the probability that, for two subjects chosen at random, the one that fails first has a worst predicted outcome. The first two VIMP measures in randomSurvivalForest estimate, using the out–of-bag (OOB) observations, how much the C–index is decreased when a covariate is perturbed. With the "random split" method, the perturbation works by randomly sending the observation to a daughter node whenever a split involving the covariate under investigation is encountered. With the "permute" method, the perturbation works by permutating the values of the covariate under investigation before sending the observations down the trees. Note that these two measures should not be interpreted as what happens if the covariate is omitted because the individual trees are still built with the covariate present. The perturbation of the covariate occurs only at prediction time. The third variable importance measure in randomSurvivalForest is based on the idea that a covariate which splits higher (closer to the top) in a tree is more important. The minimal depth used here is the average depth over the forest (0 being the smallest value and meaning a split occurring at the root node) of the highest split based on a given covariate.

Table 1 on page 62 presents a summary of the results. The "omit one covariate" measure values were already given in Figure 4 (RSF curve). We can see that the four variable importance measures tend to agree. By looking at the average rank in the last column of Table 1, we see that Bili is by far the most important covariate followed by Age and Albumin. In fact Bili is ranked in first place according to all four measures of importance. Omitting this covariate increases the IBS by 15.21%, decreases Harrell's concordance index by 0.0494 and 0.0465 for the "random split" and "permute" perturbations respectively, and the first split based on Bili occurs at depth 1.7 on average. One noticeable difference among the importance measures is that Protime ends up with more importance (2.5) according to the minimal depth measure compared to the other three.

To conclude this analysis, we provide the partial dependence plots of the two most important variables, Bili and Age. Figure 5 was produced using the

TABLE 1

Results for the four variables importance measures for the RSF model. The first one is the computer–intensive method where the models are refitted by omitting a single covariate at a time. The other three are taken from randomSurvivalForest. *Higher values indicate more importance for the first three measures and lower values indicate more importance for the minimal depth measure. The covariates are ordered according to the average rank across the four measures*

	omit one covariate	random split VIMP (× 100)	permute VIMP (× 100)	minimal depth	average rank
bili	15.21	4.94	4.65	1.7	1.00
age	3.10	0.85	1.02	2.9	2.75
albumin	3.69	−0.20	0.31	2.1	4.50
ascites	0.73	−0.06	0.08	4.4	6.63
edema	−0.01	−0.16	0.10	3.5	6.63
hepatom	−0.26	−0.02	0.10	4.8	6.63
spiders	−0.15	0.12	0.08	5.7	6.75
sex	−0.63	0.12	0.14	6.5	7.13
protime	−2.38	−0.08	0.00	2.5	7.75
drug	−0.46	−0.04	−0.12	6.1	9.00
alkphos	−2.98	−0.19	−0.02	3.5	9.50
platelet	−0.98	−0.36	−0.35	3.3	9.75

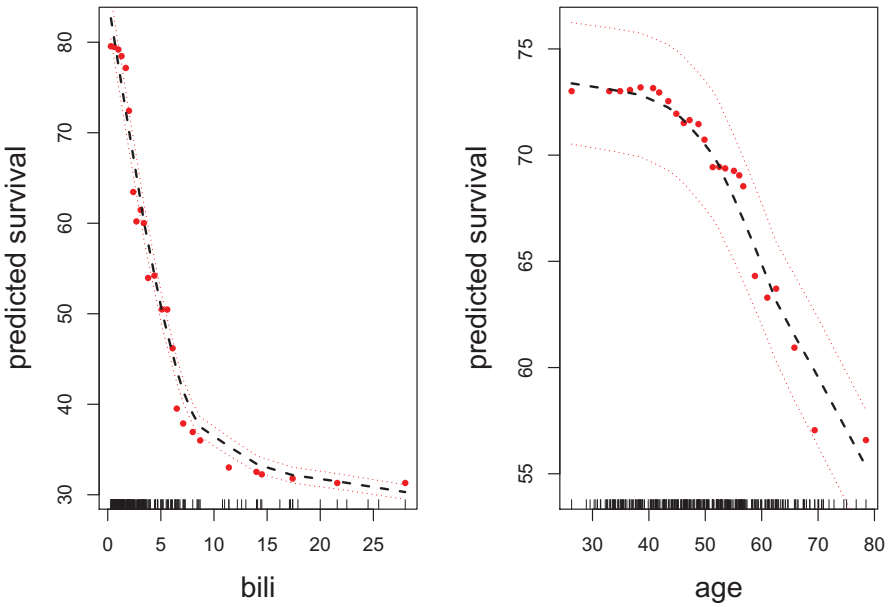

FIG 5. *Partial dependence plots for Bili and Age for the RSF model.*

plot.variable function in the randomSurvivalForest package. The y-axis is the partial predicted survival at the median follow up time which is 3395 days in our sample. More precisely, if $\hat{S}(x, x_o)$ denote the forest predicted survival at Bili=x and at x_o for the other variables (other than Bili), the partial predicted survival

at a specific value x for Bili is computed as

$$\frac{1}{N}\sum_{i=1}^{N}\hat{S}(x, x_{i,o})$$

where $x_{i,o}$ is the observed values of the other variables for observation i. Hence, it is an average (over the sample) effect. We can see that the average probability of surviving at least 3395 days decreases rapidly as Bili increases from 0 to about 8 and then the effect stabilizes somewhat. As for Bili, the probability of surviving at least 3395 days decreases with Age and the largest effect occurs after 50 years.

It is clear that other methods (tree–based and not) are available but the goal here is not to provide an exhaustive comparison across all methods but to simply illustrate some key features of survival trees using readily available implementations. The R code we used for this example is available upon request from the second author.

6. Conclusion

This review shows that survival trees have been and are still a very active area of research. Many methods have been proposed over the last 25 years. At first, the research focused on the extension of classical trees to the case of censored data. But recently, more complex models and situations have been explored and the development of ensemble methods has renewed interest in tree based methods in general and survival trees in particular. However, there are many topics that still need further research. For instance, the modeling of time–varying covariates and time–varying effects deserves much more attention.

As seen in the data example, variable importance measures and partial dependence plots can be useful to relativize the importance of the covariates and visualize their effects. However, exploring high level interactions between covariates is still problematic for survival forests. But this is not a problem specific to survival forests. It is rather a common problem to many ensemble methods for any types of outcomes (continuous, categorical and so on). However, in the context of survival trees, a further difficulty arises when time–varying effects are included. Hence, we feel that the interpretation of covariate effects with tree ensembles in general is still mainly unsolved and should attract future research. In practice, forests should be used in conjunction with more interpretable (often parametric) models. For instance, a forest could serve as a benchmark to judge the performance of more interpretable models and could validate the simpler interpretations based on them when they are judged adequate.

In principle, most existing methods to deal with missing values in trees are also applicable with survival trees. However, there is a need for a systematic investigation of the impact of missing values with survival trees and forests. Moreover, it would be interesting to develop methods to handle missing time–varying covariates.

Finally, more work is needed to extend and investigate survival trees with other types of censoring and truncation like interval–censoring, left–censoring and left–truncation.

Appendix: Selection of a single tree

Pruning methods

The pruning approach has basically two variants: cost–complexity and split–complexity. However, the basic idea is to build a large tree T_0 and obtain a sequence of nested subtrees $\{T_0, T_1, \ldots, T_M\}$ where T_M is the root–only tree. For a given tree T, we will denote by $L(T)$ and $W(T)$ the set of terminal nodes (leaves) and interior nodes of T. For a given node h of T, we will define $R(h)$ to be the within–node risk of h which measures the impurity of the node. The classical measure of impurity for a regression tree is the residual sum of squares, with the node average acting as the prediction. With survival data, many measures of impurity can be used for $R(h)$, but the choice will usually be in accordance with the splitting criterion. For instance, Leblanc and Crowley (1992) use the deviance of the node defined by $R(h) = 2(LL_h(\text{saturated}) - LL_h(\tilde{\theta}_h))$, where $LL_h(\text{saturated})$ is the log–likelihood for the saturated model with one parameter for each observation, and $LL_h(\tilde{\theta}_h)$ is the maximized log–likelihood under their adopted model. Davis and Anderson (1989) use a risk function based on the exponential log–likelihood loss.

The cost–complexity method arises from the CART paradigm. The cost–complexity of a tree is defined as

$$R_\alpha(T) = \sum_{h \in L(T)} R(h) + \alpha |L(T)|, \tag{1}$$

where α is a nonnegative parameter which governs the tradeoff between the complexity of the tree (the number of terminal nodes) and how well it fits the data. Once the cost–complexity measure is specified, the classical pruning algorithm of CART (Breiman et al., 1984) can be used to obtain the sequence of optimally pruned subtrees. Each subtree is optimal for an interval of α values.

The other method introduced by Leblanc and Crowley (1993) defines the split–complexity of a tree by

$$G_\alpha(T) = \sum_{h \in W(T)} G(h) - \alpha |W(T)|, \tag{2}$$

where $G(h)$ is the value of the standardized splitting statistic at node h (i.e., the value of the splitting criterion for the selected split at node h). Leblanc and Crowley (1993) interpret $\sum_{h \in W(T)} G(h)$ as the total amount of prognostic structure represented by the tree. Once again, the parameter α (≥ 0) governs the tradeoff between the size of the tree and how well it fits the data. Leblanc and Crowley (1993) provide an algorithm to obtain the sequence of optimal subtrees for any value of α. The split–complexity method is also used in Fan et al. (2006); Fan, Nunn and Su (2009); Bou-Hamad et al. (2009).

Final selection among the nested sequence of subtrees

Once a nested sequence of subtrees $\{T_0, T_1, \ldots, T_M\}$ has been obtained, we still need to choose one single tree in it. Many methods are available. The most popular methods are: the test set, cross–validation, bootstrap, AIC/BIC, and graphical ("kink" in the curve or elbow).

The classical CART method uses cross–validation to estimate the parameter α in the cost–complexity measure (1) and the final tree is the one corresponding to this value in the sequence of trees (Breiman et al., 1984).

With the split–complexity measure (2), Leblanc and Crowley (1993) propose two methods. The aim of both is to obtain an honest estimate of $G(T) = \sum_{h \in W(T)} G(h)$ for each tree in the sequence of subtrees, since it is clear that the in–sample values of $G(T)$ are likely to be too large. Once these are obtained, the final tree can be selected as the one maximizing (2) by fixing a value for α. Since the null distribution of their standardized splitting statistic is asymptotically χ^2_1, Leblanc and Crowley (1993) suggest using an α value in the interval $[2, 4]$. Their argument is that $\alpha = 2$ is in the spirit of the AIC criterion while $\alpha = 4$ corresponds roughly to using a 0.05 significance level for the χ^2_1 distribution. Their first method consists in applying a bootstrap bias correction to $G(T)$ and is applicable with any sample size. Their second method is useful for large samples and consists in dividing the original sample into training and test samples. The training sample is used to build the large tree and obtain the sequence of subtrees. The test sample is then used to recompute the value of $G(T) = \sum_{h \in W(T)} G(h)$ for each tree in the sequence. The optimal tree is then chosen using the recomputed values of (2).

The AIC/BIC type methods proposed in other work are closely related to the second method of Leblanc and Crowley (1993). The selection methods proposed in Ciampi et al. (1987); Su and Fan (2004), and Su and Tsai (2005) all involve selecting the final tree, among a sequence of subtrees, as the one minimizing a criterion like

$$-2ll(T) + \alpha |L(T)|$$

where $ll(T)$ is the log–likelihood of the tree and α is either 2 (AIC) or $\log(n)$ (BIC). The whole procedure involves building a large tree and obtaining a sequence of subtrees with a training sample and then recomputing $ll(T)$ with a test sample.

Graphical methods that plot the value of a criterion as a function of the tree complexity for each tree in the sequence have also been proposed. Similar to a scree plot in a principal components analysis, such a plot usually has an elbow shape with an abrupt change at some point. The final tree is then the one corresponding to the "kink" in the curve. Segal (1988) proposes such a method coupled with a specific pruning method. For this method, each internal node is assigned the maximum split statistic in the subtree of which the node is the root. This method is also used in Gao, Manatunga and Chen (2004). One drawback of graphical methods is the subjectivity associated with them. Negassa et al. (2000) propose an automatic elbow detection method and apply it with an AIC

criterion (as above) but compute it on the same sample as the one that built the tree.

Forward methods

When the covariates are measured on different scales, the number of candidate splits at a given node can be very different for each covariate. For instance, if the splitting criterion is based on a p–value, then a covariate with a higher number of tests has a greater probability of achieving a small p–value. This is why the use of adjusted p–values has been proposed to avoid any possible selection bias in the choice of the covariate (Schlittgen, 1999; Lausen et al., 2004).

At the same time, adjusted p–value can be used to regulate the tree–building procedure, acting as a criterion for when to stop splitting a node further. Using such a rule gives rise to a forward method which avoids the use of pruning. Using the standardized two–sample logrank statistic as the splitting criterion, Lausen et al. (2004) propose such a method Their method not only accounts for the fact that multiple tests are performed for each covariate but also for the fact that many covariates are involved, and hence that the overall best value of the test statistic is a maximum (over the covariates) of maximally selected statistics (over all potential splits on a covariate). Splitting is stopped when the adjusted p–value of the selected best split is greater than a pre–specified value (for instance 0.05).

Acknowledgement

The authors would like to thank two reviewers and an Associate Editor for their comments that helped us prepare a more complete, readable and appealing version of the paper. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

References

Ahn, H. and Loh, W.-Y. (1994). Tree–Structured Proportional Hazards Regression Modeling. *Biometrics* **50**, 471–485.

Bacchetti, P. and Segal, M. (1995). Survival Trees with Time–Dependent Covariates: Application to Estimating Changes in the Incubation Period of AIDS. *Lifetime Data Analysis* **1**, 35–47.

Benner, A. (2002). Application of "Aggregated Classifiers" in Survival Time Studies. *COMPSTAT 2002 - Proceedings in Computational Statistics: 15th Symposium Held in Berlin, Germany, 2002* MR1973489

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., Mâsse, L., Vitaro, F. and Tremblay, R. (2009). Discrete–Time Survival Trees. *Canadian Journal of Statistics* **37**, 17–32. MR2509459

Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2011). Discrete–Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data. To appear in *Statistical Modeling*.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California. MR0726392

Breiman, L. (1996). Bagging Predictors. *Machine Learning* **24**, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5–32.

Cho, H. and Hong, S-M. (2008). Median Regression Tree for Analysis of Censored Survival Data. *Systems, Man and Cybernetics, Part A, IEEE Transactions on* **38**, 715–726.

Ciampi, A., Bush, R. S., Gospodarowicz, M. and Till, J. E. (1981). An Approach to Classifying Prognostic Factors Related to Survival Experience for Non–Hodgkin's Lymphoma Patients: Based on a Series of 982 Patients: 1967–1975. *Cancer* **47**, 621–627.

Ciampi, A., Thiffault, J., Nakache, J.-P. and Asselain, B. (1986). Stratification by Stepwise Regression, Correspondance Analysis and Recursive Partition: A Comparison of Three Methods of Analysis for Survival Data with Covariates. *Computational Statistics & Data Analysis* **4**, 185–204.

Ciampi, A., Chang, C. H., Hogg, S. and McKinney, S. (1987). Recursive Partition: A Versatile Method for Exploratory Data Analysis in Biostatistics. *Biostatistics* 23–50

Ciampi, A., Hogg, S. A., McKinney, S. and Thiffault, J. (1988). RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics. I. Methods and Program Features. *Computer Methods and Programs in Biomedicine* **26**, 239–256.

Ciampi, A., Hogg, S. A., McKinney, S. and Thiffault, J. (1989). RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics. II. Applications to Data on Small Cell Carcinoma of the Lung (SCCL). *Computer Methods and Programs in Biomedicine* **30**, 283–296.

Ciampi, A., Negassa, A. and Lou, Z. (1995). Tree–Structured Prediction for Censored Survival Data and the Cox Model. *Journal of Clinical Epidemiology* **48**, 675–689.

Ciampi, A., Thiffault, J. and Sagman, U. (1989). RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics. II. Applications to Data on Small Cell Carcinoma of the Lung (SCCL). *Computer Methods and Programs in Biomedicine* **30**, 283–296.

Dannegger, F. (2000). Tree Stability Diagnostics and Some Remedies for Instability. *Statistics in Medicine* **19**, 475–491.

Davis, R. B. and Anderson, J. R. (1989). Exponential Survival Trees. *Statistics in Medicine* **8**, 947–961.

Ding, Y. and Simonoff, J. S. (2010). An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data. *Journal of Machine Learning Research*, **11**, 131–170. MR2591624

Eckel, K. T., Pfahlberg, A., Gefeller, O. and Hothorn, T. (2008). Flexible Modeling of Malignant Melanoma Survival. *Methods of Information in Medicine* **47**, 47–55.

Fan, J., Nunn, M. E. and Su, X. (2009). Multivariate Exponential Survival Trees and Their Application to Tooth Prognosis. *Computational Statistics and Data Analysis* **53**, 1110–1121. MR2657075

Fan, J., Su, X.-G., Levine, R., Nunn, M. and Leblanc, M. (2006). Trees for Censored Survival Data by Goodness of Split, with Application to Tooth Prognosis. *Journal of American Statistical Association* **101**, 959–967. MR2324107

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New Jersey. MR1100924

Gao, F., Manatunga, A. K. and Chen, S. (2004). Identification of Prognostic Factors with Multivariate Survival Data. *Computational Statistics & Data Analysis* **45**, 813–824. MR2054888

Gao, F., Manatunga, A. K. and Chen, S. (2006). Developing Multivariate Survival Trees with a Proportonal Hazards Structure. *Journal of Data Science* **4**, 343–356.

Gordon, L. and Olshen, R. A. (1985). Tree–structured Survival Analysis. *Cancer Treatment Reports* **69**, 1065–1069.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and Comparisons of Prognostic Classification Schemes for Survival Data. *Statistics in Medicine* **18**, 2529–2545.

Hammer, P. L. and Bonates, T. O. (2006). Logical Analysis of Data–An Overview: From Combinatorial Optimization to Medical Applications. *Annals of Operations Research* **148**, 203–225.

Harrell, F., Califf, R., Pryor, D., Lee, K. and Rosati, R. (1982). Evaluating the Yield of Medical Tests. *Journal of the American Medical Association* **247**, 2543–2546.

Hothorn, T., Lausen, B., Benner, A. and Radespiel–Tröger, M. (2004). Bagging Survival Trees. *Statistics in Medicine* **23**, 77–91.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. M. and van der Laan, M. J. (2006). Survival Ensembles. *Biostatistics* **7**, 355–373.

Huang, X., Chen, S. and Soong, S. (1998). Piecewise Exponential Survival Trees with Time–Dependent Covariates. *Biometrics*, **54**, 1420–1433.

Ishwaran, H., Blackstone, E. H., Pothier, C. E. and Lauer, M. S. (2004). Relative Risk Forests for Exercise Heart Rate Recovery as a Predictor of Mortality. *Journal of the American Statistical Association* **99**, 591–600. MR2086385

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random Survival Forests. *Annals of Applied Statistics* **2**, 841–860. MR2516796

Ishwaran, H. and Kogalur, U. B. (2010a). Consistency of Random Survival Forests. *Statistics and Probability Letters* **80**, 1056–1064. MR2651045

Ishwaran, H. and Kogalur, U. B. (2010b). Random Survival Forests, R package version 3.6.3.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. and Lauer, M. S. (2010). High Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* **105**, 205–217.

Jin, H., Lu, Y., Stone, K. and Black, D. M. (2004). Alternative Tree–Structured Survival Analysis Based on Variance of Survival Time. *Medical Decision Making* **24**, 670–680.

Keles, S. and Segal, M. R. (2002). Residual–Based Tree–Structured Survival Analysis. *Statistics in Medicine* **21**, 313–326.

Krętowska, M. (2004). Dipolar Regression Trees in Survival Analysis. *Biocybernetics and Biomedical Engineering* **24**, 25–33.

Krętowska, M. (2006). Random Forests of Dipolar Trees for Survival Prediction. Artificial Intelligence and Soft Computing - ICAISC 2006, Proceedings. *Lecture Notes In Computer Science* **4029**, 909–918.

Krętowska, M. (2010). The influence of Censoring for the Performance of Survival Tree Ensemble. Artificial Intelligence and Soft Computing, Pt II - ICAISC 2010, Proceedings. *Lecture Notes in Artificial Intelligence* **6114**, 524-531.

Kronek, L. P., and Reddy, A. (2008). Logical Analysis of Survival Data: Prognostic Survival Models by Detecting High–Degree Interactions in Right–Censored Data. *Bioinformatics* **24**, 248–253.

Lausen, B., Hothorn, T., Bretz, F. and Schumacher, M. (2004). Assessment of Optimal Selected Prognostic Factors. *Biometrical Journal* **46**, 364–374. MR2079857

LeBlanc, M. and Crowley, J. (1992). Relative Risk Trees for Censored Survival Data. *Biometrics* **48**, 411–425.

LeBlanc, M. and Crowley, J. (1993). Survival Trees by Goodness of Split. *Journal of the American Statistical Association* **88**, 457–467. MR1224370

LeBlanc, M. and Crowley, J. (1995). A Review of Tree–Based Prognostic Models. *Journal of Cancer Treatment and Research* **75**, 113–124.

Loh, W-Y. (1991). Survival Modeling Through Recursive Stratification. *Computational Statistics and Data Analysis* **12**, 295–313.

Marubini, E., Morabito, A. and Valsecchi, M. G. (1983). Prognostic Factors and Risk Groups: Some Results Given by Using an Algorithm Suitable for Censored Survival Data. *Statistics in Medicine* **2**, 295–303.

Molinaro, A. M., Dudoit, S. and van der Laan, M. J. (2004). Tree–based Multivariate Regression and Density Estimation with Right–censored Data. *Journal of Multivariate Analysis* **90**, 154–177. MR2064940

Morgan, J. and Sonquist, J. (1963). Problems in the Analysis of Survey Data and a Proposal. *Journal of the American Statistical Association* **58**, 415–434.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. and Boivin, J.-F. (2000). Tree–Structured Prognostic Classification for Censored Survival Data: Validation of Computationally Inexpensive Model Selection Criteria. *Journal of Statistical Computation and Simulation* **67**, 289–318. MR1815167

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. and Boivin, J.-F. (2005) Tree–Structured Subgroup Analysis for Censored Survival Data: Validation of Computationally Inexpensive Model Selection Criteria. *Statistics and Computing* **15**, 231–239. MR2147555

Peters, A. and Hothorn, T. (2009). ipred: Improved Predictors. R package version 0.8-8. http://CRAN.R-project.org/package=ipred.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Radespiel-Tröger, M., Rabenstein, T., Schneider, H. T. and Lausen, B. (2003). Comparison of Tree–based Methods for Prognostic Stratification of Survival Data. *Artificial Intelligence in Medicine* **28**, 323–341.

Radespiel-Tröger, M., Gefeller, O., Rabenstein, T. and Hothorn, T. (2006). Association Between Split Selection Instability and Predictive Error in Survival Trees. *Methods of Information in Medicine* **45**, 548–556.

Ridgeway, G. (1999). The State of Boosting. *Computing Science and Statistics*. **31**, 172–181.

Rokach, L. (2008). Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography. *Computational Statistics and Data Analysis* **53**, 4046–4072. MR2744304

Schlittgen, R. (1999). Regression Trees for Survival Data – an Approach to Select Discontinuous Split Points by Rank Statistics. *Biometrical Journal* **41**, 943–954. MR1747521

Segal, M. R. (1988). Regression Trees for Censored Data. *Biometrics* **44**, 35–48.

Segal, M. R. (1992). Tree–Structured Methods for Longitudinal Data. *Journal of the American Statistical Association* **87**, 407–418.

Siroky, D.S. (2009). Navigating Random Forests and Related Advances in Algorithmic Modeling. *Statistics Surveys* **3**, 147–163. MR2556872

Su, X. and Fan, J. (2004). Multivariate Survival Trees: A Maximum Likelihood Approach Based on Frailty Models. *Biometrics* **60**, 93–99. MR2043623

Su, X. and Tsai, C.-L. (2005). Tree–augmented Cox Proportional Hazards Models. *Biostatistics* **6**, 486–499.

Therneau, T., Grambsch, P. and Fleming, T. (1990). Martingale–Based Residuals for Survival Models. *Biometrika* **77**, 147–160. MR1049416

Therneau, T. M. and Atkinson, B. (2010). R port by Brian Ripley. rpart: Recursive Partitioning. R package version 3.1-46. http://CRAN.R-project.org/package=rpart.

Tsai, C., Chen, D.-T., Chen, J., Balch, C. M., Thompson, J. and Soong, S.-J. (2007). An Integrated Tree–Based Classification Approach to Prognostic Grouping with Application to Localized Melanoma Patients. *Journal of Biopharmaceutical Statistics* **17**, 445–460. MR2370755

Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011). Mining Data With Random Forests: A Survey and Results of New Tests. *Pattern Recognition* **44**, 330–349.

Wallace, M. L., Anderson, S. J. and Mazumdar, S. (2010). A Stochastic Multiple Imputation Algorithm for Missing Covariate Data in Tree–Structured Survival Analysis. *Statistics in Medicine* **29**, 3004–3016.

Xu, R. and Adak, S. (2001). Survival Analysis with Time–Varying Relative Risks: A Tree–Based Approach. *Methods of information in medicine* **40**, 141–147.

Xu, R. and Adak, S. (2002). Survival Analysis with Time–Varying Regression Effects Using a Tree–Based Approach. *Biometrics* **58**, 305–315. MR1908170

Yin, Y. and Anderson, J. (2002). Nonparametric Tree-Structured Modeling for Interval-Censored Survival Data. Joint Statistical Meeting, August 2002. 6 pages. MR2703473

Zhang, H.P. (1995). Splitting Criteria in Survival Trees. *In Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modeling*, 305–314, Springer.