Discrete-Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data

Article *in* Statistical Modelling · October 2011

DOI: 10.1177/1471082X1001100503

CITATIONS
35

READS
1,987

3 authors, including:

Imad Bou--Hamad
American University of Beirut
26 PUBLICATIONS **509** CITATIONS

SEE PROFILE

Les Cahiers du GERAD ISSN: 0711–2440

Discrete-Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data

I. Bou-Hamad, D. Larocque, H. Ben-Ameur

G–2009–17

March 2009

Discrete-Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data

Imad Bou-Hamad
Denis Larocque
Hatem Ben-Ameur

GERAD and Department of Management Sciences
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7

imad.bou-hamad@hec.ca
denis.larocque@hec.ca
hatem.ben-ameur@hec.ca

March 2009

Les Cahiers du GERAD
G–2009–17

Abstract

Discrete-time survival data with time-varying covariates are often encountered in practice. One such example is bankruptcy studies where the status of each firm is available on a yearly basis. Moreover, these studies often use financial and accounting based ratios to predict bankruptcy. These ratios are also yearly measures and hence are time-varying. In this paper, we propose a new survival tree method for discrete-time survival data with time-varying covariates. This method can accommodate simultaneously time-varying covariates and time-varying effects. The new method is applied to a sample of United States firms that conducted an Initial Public Offerings between 1990 and 1999.

Key Words: Bankruptcy data; discrete-time survival analysis; survival forests; time-varying covariate.

Résumé

Des données de survie à temps discret sont souvent présentes en pratique. Un tel exemple est l'étude de faillites où le statut des firmes est connu annuellement. De plus, ces études utilisent souvent des ratios comptables ou financiers pour prédire la faillite. Ces ratios sont aussi des mesures annuelles et donc ils varient dans le temps. Dans cet article, nous proposons une nouvelle méthode pour construire des arbres de survie à temps discret avec des covariables qui varient dans le temps. Cette méthode peut à la fois contenir des covariables qui varient dans le temps et des effets qui varient dans le temps. La méthode est appliquée à un échantillon de firmes américaines qui ont fait un premier appel public à l'épargne entre 1990 et 1999.

1 Introduction

The analysis and prediction of corporate financial distress and bankruptcy are important problems that generated many theoretical and empirical research over the last four decades. The use of financial and accounting based ratios to predict bankruptcy goes back to Beaver (1966). Since then, many modeling techniques using these ratios have been proposed. Some popular ones are the multivariate discriminant analysis (Altman, 1968), linear regression (Meyer and Pifer, 1970), logistic regression (Ohlson, 1980), probit model (Zmijewski, 1984), classification tree (Frydman, Altman and Kao, 1985) and neural network (Fanning and Cogger, 1994). However, the methods above do not take in consideration the change of firms characteristics over time and hence are called static or single-period models. More precisely, only one set of covariates recorded at a single period in time is used to model bankruptcy at a fixed moment in the future (usually between one to three years later). Since bankruptcies are rare events, samples are usually collected over a a long period. Consequently, several years of data are available on the firms of interest. By using only the covariates at a single period, static models waste a lot of information.

Shumway (2001) handled the problem of change through time using a discrete-time hazard model that allows the use of many years of data for each firm. A discrete-time approach is appropriate since the usual covariates (ratios) and the bankruptcy indicator are yearly measures. Moreover, the discrete-time approach can easily incorporate time-varying covariates. This approach has been extended since then. For instance, De Leonardis and Rocci (2008) proposed a discrete-time survival model with frailty to allow for unobserved heterogeneity and Nam, Kim, Park and Lee (2008) incorporated macroeconomic dependencies. These studies showed the benefits of a multiple-period approach over a single-period approach since they report better predictive accuracies. However, they are all based more or less on the same parametric logit or log-log formulation for the hazard function. Other approaches might produce better results. Survival trees is another modeling strategy that will be investigated in this paper.

Tree-based methods (Morgan and Sonquist, 1963, Breiman, Friedman, Olshen and Stone, 1984), and survival trees (Gordon and Olshen, 1985) in particular are now well established techniques that are popular among practitioners. Even though a single tree is often a very useful descriptive and predictive tool in itself, the development of ensemble methods like bagging (Breiman, 1996), and random forests (Breiman, 2001) unleashed all their potential predictive power when a tree is used as the base learner. Many survival tree methods were proposed in the last twenty years. Some use the log-rank statistic as a splitting criterion (Ciampi, Thiffault, Nakache and Asselain, 1986, Segal, 1988, and LeBlanc and Crowley, 1993), while others use likelihood approaches to derive a splitting criterion (Davis and Anderson, 1989, LeBlanc and Crowley, 1992). Other methods include Molinaro, Dudoit and Van Der Laan, (2004), Jin, Lu, Stone and Black (2004) and Su and Tsai (2005). Recently, extensions to multivariate and correlated data were proposed (Su and Fan, 2004, Gao, Manatunga and Chen, 2004, and Fan, Su, Levine, Nunn and LeBlanc, 2006). Finally, ensemble methods applied to survival trees were studied in Hothorn, Lausen, Benner and Radespiel-Tröger (2004), Hothorn, Bühlmann, Dudoit, Molinaro and Van Der Laan (2006) and Ishwaran, Kogalur, Blackstone and Lauer (2008).

These methods were mainly developed for continuous-time survival variables. A method designed for discrete-time variables was proposed in Bou-Hamad, Larocque, Ben-Ameur, Mâsse, Vitaro and Tremblay (2009). However, only time independent covariates can be incorporated with this approach.

Since time-varying covariates are common in practice, it is surprising that very little work has been devoted to the topic of extending survival trees to allow them to incorporate such covariates. Segal (1992) underlined that no convincing technique for defining splits on time-varying covariates has been developed. The only strategy that had been implemented at that time consisted in replacing the time-varying covariate with a low–order polynomial approximation. In particular, linear summaries have been used where each time-varying covariate is first regressed against time within individuals. The intercept and slope for each individual are then used as covariates. Obviously, such a method is only reasonable when the linear regression adequately captures the time-varying covariate, but a serious limitation arises when the number of observations per subject is small since the intra-individual regressions will be imprecise. Later, Bacchetti and Segal (1995)

proposed to handle time-varying covariates by decomposing each subject survival experience into pseudo-subjects according to the values of the splitting rules. More precisely, when considering to split a node of the tree, a subject can be splitted apart across the two children nodes. The time window where the splitting rule is true would go to one node (say the left node), and the time window where it is false would go to the other node (say the right node). A discrete version of this method is basically the one retained in this paper, and will be discussed into more details in Section 2.2. Finally, Huang, Chen and Soong (1998) proposed a piecewise exponential survival tree method that accommodates time-varying covariates. This method assumes that the distribution of the survival time for a subject is given by a piecewise exponential distribution with k pieces and, as the Bacchetti and Segal (1995) method, allows subjects to be splitted across different nodes. However, these developments are aimed at continuous-time survival data.

The goal of this paper is to extend the discrete-time survival tree method introduced in Bou-Hamad et al. (2009) to the case of time-varying covariates and provide an application to bankruptcy data.

The rest of the paper is organized as follows. In Section 2, we describe the basic tree building method and show how single trees can be combined to form a survival forest. The bankruptcy data application is detailed in Section 3. Section 4 presents concluding remarks.

2 Description of the tree building method

The proposed tree method is an extension of the one introduced in Bou-Hamad et al. (2009) and is built around a discrete-time proportional odds (DTPO) model that was popularized by Singer and Willett (1993).

2.1 Data description and discrete-time hazard models

Data on N independent subjects are available. For subject i, we observe $(\tau_i, \delta_i, \mathbf{x}_i)$ consisting of 1) a discrete survival time τ_i $(\in \{1, 2, \ldots\})$, 2) a censoring indicator δ_i taking a value of 1 if the true time-to-event is observed and 0 if it is right-censored, and 3) a set of values for p covariates $\mathbf{x}_i$. Some covariates can be time independent and some others can be time-varying. We will denote by $x_{ki}(j)$ the value of the k^{th} covariate at time j for subject i. Even thought we use the same notation for all covariates, it is clear that $x_{ki}(j)$ remains constant for all j for a time independent covariate. Denoting by U_i the real time-to-event for subject i, which is unobserved for the censored subjects, and suppressing the dependence on the covariates to simplify the notation, we define

$$h_i(j) = P(U_i = j | U_i \geq j), \quad S_i(j) = P(U_i > j), \quad \text{and} \quad \pi_i(j) = P(U_i = j) \tag{1}$$

as the hazards, the survival probabilities and the probabilities of events, respectively. We also make the usual assumption that U_i and the true censoring time are independent given the covariates.

Assume that K is the maximum observed time for the data. The basic DTPO model, as described in Singer and Willett (1993), is

$$\log\left(\frac{h_i(j)}{1-h_i(j)}\right) = \alpha_1 D_{1i}(j) + \cdots + \alpha_K D_{Ki}(j) + \beta_1 x_{1i}(j) + \ldots + \beta_p x_{pi}(j), \tag{2}$$

where the $D_{ki}(j)$'s are indicator variables indexing the time periods that are defined by $D_{ki}(j) = 1$ if $k = j$ and 0 otherwise. Fitting this model by maximum likelihood is easy when we realize that the likelihood function of (2) is equivalent to an independent Bernoulli trials model with transformed data with a logistic dependence on the covariates (see page 171 of Singer and Willett, 1993). Hence, any logistic regression software can be used to fit this model. Moreover, the proportional odds assumption can be easily relaxed by introducing interaction terms between a covariate and the time indexing indicators. For instance, with only one covariate C, which may be time-varying, the resulting model would be

$$\log\left(\frac{h_i(j)}{1-h_i(j)}\right) = \alpha_1 D_{1i}(j) + \cdots + \alpha_K D_{Ki}(j) + \beta_1 C_i(j) D_{1i}(j) + \ldots + \beta_K C_i(j) D_{Ki}(j). \tag{3}$$

This model is important because, as we will see in the next subsection, it is the one we use to derive the splitting criterion for the tree building algorithm. With this model, not only can a time-varying covariate be used, but its effect is also allowed to be time dependent.

Note that the baseline effect of time is modeled in the most flexible way in (2) since each time period has its own parameter. It is possible to simplify the model and specify a linear or constant time effect. For instance, Shumway (2001) used a constant time effect in his bankruptcy model.

In general, the log-likelihood function of a discrete-time survival model can be written as

$$LL = \sum_{i=1}^{N} \delta_i \ln(\pi_i(\tau_i)) + (1 - \delta_i) \ln(S_i(\tau_i)). \tag{4}$$

Moreover, the maximum likelihood estimates (MLEs) of the hazards in model (2) but without covariates, that is the model $\log(h_i(j)/(1 - h_i(j))) = \alpha_1 D_{1i}(j) + \cdots + \alpha_K D_{Ki}(j)$, are given by

$$\hat{h}(j) = \frac{e(j)}{r(j)} \quad \text{for} \quad j = 1, \ldots, K, \tag{5}$$

where $e(j)$ is the number of subjects that experienced the event at time j, and $r(j)$ is the number of subjects that were at risk at time j. Defining $\hat{S}(0) = 1$, the MLEs of the survival and probability functions are then obtained as $\hat{S}(j) = \hat{S}(j-1)(1 - \hat{h}(j))$ and $\hat{\pi}(j) = \hat{S}(j-1) - \hat{S}(j)$.

2.2 Tree building

We assume the reader is familiar with the basic terminology used with tree based methods (Breiman et al., 1984). The first important aspect concerning tree building is the splitting criterion. This criterion will be used to partition the sample according to binary rules based on the covariates. If a single tree is needed, the usual procedure consists in building a large tree, to prune some branches off and to select one tree among a nested sequence of pruned trees as will be described in this subsection. Another strategy is to use trees as the basic model in an ensemble method like bagging and random forests. With this strategy, many trees are built (usually without pruning) and combined as described in the next subsection.

Let x be any covariate (time-varying or not). If x is continuous or ar least ordinal, any splitting variable will have the form $C_i(j) = I(x_i(j) \leq c)$ where I is the indicator function and $x_i(j)$ is the value of x at time j for subject i. For a categorical covariate, any splitting variable will have the form $C_i(j) = I(x_i(j) \in \{c_1, \ldots, c_l\})$ where $c_1, \ldots, c_l$ are possible values of x. For the retained splitting variable, the observations for which $C_i(j) = 1$ would go to the right node while the ones for which $C_i(j) = 0$ would go to the left node. Note that we are now using the word "observation" and not "subject". This is because we must now shift to a "subject-period" data set point of view (Singer and Willett, 1993) where each subject has one line of observation for each period where he is at risk. Usually, this means that a subject has one line of observation for each period until he experiences the event or is censored. If the splitting variable is defined through a time independent variable, then the condition is either true or false for all periods. Hence, all the observations (lines in the subject-period data set) for this subject would go to the same node. However, if the splitting variable is defined through a time-varying variable, it is possible that the condition is true for some periods and false for the others. Hence, some observations could go to one node and some others could go to the other node which means that the subject could be splitted across the two children nodes.

The splitting criterion we are proposing is based on the observed log-likelihood of model (3) where the C variable is now a splitting variable as above. Since C is an indicator variable, the contribution to the total likelihood of the observations for which $C = 1$ is separated from the contribution of the observations for which $C = 0$. Hence, fitting the model amounts to fit two separate models, one using the observations for which $C = 1$ (right node) and one for the others (left node). But these are intercepts only models (one parameter for each time period) and the MLE's of the hazards, survival function and probability function are given by (5) and below. Note that we are only using the observations that are in the right (left) node to

compute the MLEs of the right (left) node. The splitting criterion is then given by (4) by plugging in the values of the MLEs.

The tree building can now be described as follows. Start with all observations (all lines in the subject-period data set) in the root node. Compute the value of the splitting criterion (observed LL) for all possible splitting variables constructed with all possible covariates. The optimal split is the one with the maximum value for the splitting criterion. Using the optimal splitting variable, split the observations across the two children nodes. Split the right node with the same procedure using only the observations in the node and do the same for the left node. Repeat the process recursively until a stopping criterion is reached. For instance, do not split a node further when it contains less than a predetermined number of observations.

It is clear that in the end, any given subject can be splitted across many nodes. This is also happening with the method proposed in Bacchetti and Segal (1995). However, their approach was aimed at a continuous time survival variable and the effect of the splitting variable remained time-invariant. With our method the effect of the splitting variable depends on the period due to the interactions between this variable and the time indicators. Hence, our method imposes less assumptions. As a side effect, it also allows a closed form expression for the splitting criterion and thus speeds up the computations which is important when the number of observations and covariates are large. However, more parameters need to be estimated and this can become impractical when the number of periods is large. But in this case, it would be possible to treat the survival variable as a continuous one and use one the many available survival tree methods for continuous data. Another possibility is to base the splitting criterion on a restricted model. Model (3) used in this paper is basically the most general model. At the other extreme, the simplest one would be

$$\log\left(\frac{h_i(j)}{1-h_i(j)}\right) = \beta_0 + \beta_1 C_i(j) \tag{6}$$

with time-independent effects for both the period and the splitting variable. Any intermediate model between the two extreme ones (3) and (6) are also possible. But these models would require numerical computations of the MLE's and computation time could become an issue. When the data contains only time independent covariates, the splitting criterion reduces to the one of Bou-Hamad et al. (2009). Hence the method proposed here is a direct extension of the earlier method that allows the use of time-varying covariates.

If a single tree must be chosen, we are proposing to use the same pruning and selection method as in Bou-Hamad et al. (2009). It is basically based on the split complexity measure of LeBlanc and Crowley (1993) combined with the bootstrap. The reader is referred to Section 2.3 of Bou-Hamad et al. (2009) for more details.

2.3 Bagging and survival forests

It is now well-known that averaging many trees through an ensemble method produces often a better model than a single tree; Breiman (1996, 2001) and Hamza and Larocque (2005) for classification and regression trees and Hothorn et al. (2004), Hothorn et al. (2006) and Ishwaran et al. (2008) for survival trees. Bagging was studied in Bou-Hamad et al. (2009). In this paper, we will present the slightly more general concept of a survival forest.

The general method goes as follows: 1) Draw B bootstrap samples from the original data, 2) Grow a tree for each bootstrap sample. At each node, select at random k out of p covariates where $k \in \{1, \ldots, p\}$ is a parameter chosen by the analyst at the start. No pruning is performed. The splitting is stopped when a minimum node size is reached.

To obtain estimated hazards, survival probabilities and probabilities for the i^{th} observation of the data to be scored:

1. Let the observation fall into each tree. Let $\hat{\pi}_i^b(j)$ denote the estimate of $\pi_i(j)$, $j = 1, \ldots, K$, obtained from the b^{th} tree, $b = 1, \ldots, B$.

2. Let $\hat{\pi}_i(j) = \frac{1}{B}\sum_{b=1}^{B} \hat{\pi}_i^b(j)$ denote the final ensemble estimate of $\pi_i(j)$.
3. The ensemble estimate of the survival probabilities are then calculated recursively as $\hat{S}_i(j) = \hat{S}_i(j-1) - \hat{\pi}_i(j)$ and $\hat{h}_i(j) = \hat{\pi}_i(j)/\hat{S}_i(j-1)$ for $j = 1, \ldots, K$, where $\hat{S}_i(0) = 1$. Note that $\hat{h}_i(j)$ is defined to be 0 when $\hat{S}_i(j-1) = 0$.

Selecting all the covariates in each node amounts to perform bagging which is a particular case of the random (survival) forests.

3 Application to bankruptcy data

3.1 Description of the data

Our study focuses on United States firms that conducted IPOs (Initial Public Offerings) between 1990 and 1999. IPOs are often used by smaller, younger companies seeking the capital to expand, but can also be done by large privately owned companies looking to become publicly traded. IPOs were the most prevalent form of securities issued to raise capital in the United States in the last decade (1990-2000). The Sample was collected from the COMPUSTAT database. The target variable is bankruptcy. All firms that filed for bankruptcy under Chapter 7 or 11 are considered bankrupt. The covariates are financial ratios. Since there is a substantial quantity of accounting statements, there is a huge number of ratios that can be calculated. However, financial ratios are usually grouped into five categories (Ross, Westerfield, Jordan and Roberts, 2002, Chapter 3, Section 3.3): 1) short-term solvency or liquidity ratios, 2) turnover or activity ratios, 3) financial leverage or long-term solvency ratios, 4) profitability ratios, and 5) market value ratios. One ratio has been selected from each class to represent it. The candidate ratio is the one which was mostly used in previous studies as indicated in the review paper by Bellovary, Giacomino and Akers (2007). The selected ratios are:

- R_1 = Current Assets/Current Liabilities
- R_2 = Sales/Total Assets
- R_3 = Total Debt/Total Assets
- R_4 = Net Income/Total Assets
- R_5 = Market Value of Equity/Book Value of Total Debt.

Each firm is followed yearly starting from its initial IPO until 2004. Hence, the ratios are available on a yearly basis, and are treated as time-varying covariates. In order to make the modeling exercise realistic, the ratios are used to model the bankruptcy indicator (1=yes, 0=no) at an horizon of three years. Hence, we are trying to relate the values of the ratios in a given year to the bankruptcy indicator three years later. The sample has 1143 firms, 189 of them went bankrupt during the study period. However, since 174 of the 189 bankruptcies occurred between years 3 and 8 after the IPO, only this six years period is retained for the final analysis. The 15 remaining bankruptcies, which are scattered among the eight remaining years, do not convey enough information to allow accurate estimations. In the end, the data set contains 6202 firm-year observations. The tree building algorithm is implemented in Ox (Doornik, 2002), and the maximum likelihood estimation of the DTPO model is implemented in R (R Development Core Team, 2007).

Table 1 presents the empirical risks for the six periods under study. Note that the first line of the table (3 years after the IPO) includes the firms that went bankrupt in years 1, 2 or 3 after the IPO.

Table 2 presents summary statistics for the five retained covariates (ratios). The distributions of the ratios are skewed, especially for R_1 and R_5, and this is why various transformations were performed on them as described below.

Table 1: Empirical risks for the bankruptcy data.

Year after IPO	Number of firms at risk	Number of bankruptcies	Risk (%)
3	1143	35	3.06
4	1108	41	3.70
5	1067	34	3.19
6	1033	29	2.81
7	1004	18	1.79
8	986	17	2.01

Table 2: Summary statistics for the ratios (n=6202).

Ratio	Min	Max	Mean	Median	Std
R_1	0.046	258.27	3.84	2.29	6.26
R_2	0.000	15.96	1.09	0.92	0.98
R_3	0.005	9.34	0.47	0.42	0.44
R_4	-23.99	1.69	-0.10	0.03	0.64
R_5	0.000	749.84	14.15	3.71	37.46

3.2 Results

Three types of models are fitted to the data and compared: 1) DTPO models, 2) single trees and, 3) survival forests.

The parameter estimates of some DTPO models are presented in Table 3. The basic model using the original ratios is in the second column, but according to the AIC and BIC criteria, this model is inferior to the other three models that use transformed ratios. To alleviate the skewness effect, the first transformation uses truncated ratios (third column in Table 3). The ratios were truncated at their 95% quantile, i.e., any value above the quantile was brought back down to the quantile value. The ratio R_4 was also truncated above its 5% quantile value because it is also skewed to the left. In another model (fourth column in Table 3), the transformation $\log(R_i + 2)$ for $i = 1, 2, 3, 5$ and $\log(-R_4 + 2)$ were used. However, according to the AIC and BIC criteria, the best result was obtained for what we call the "MAD-log" transformation (last column in Table 3). This transformation is defined as follows. First we standardize the ratio by subtracting the median and dividing by the MAD (mean absolute deviations), which are highly robust location and scale measures. Then we apply the transformation $\text{sign}(x)\log(|x| + 1)$ to the standardized data. As for the the other two transformations, the MAD-log transformation is monotonic. Hence the sign of the effects are comparable across all models. For the MAD-log model, only R_4 and R_5 are significant with negative effects. Thus, higher risks of bankruptcy are associated with lower values of R_4 (Net Income/Total Assets) and R_5 (Market Value of Equity/Book Value of Total Debt). We also investigated if a year effect was present by incorporating the year of the IPO as a covariate but it turns out to be non-significant in all models.

The proposed tree method was then applied to the data. The tree presented in Figure 1 is the one obtained after pruning and selecting the best one with 30 bootstrap samples. The number of observations, the number of bankruptcies and the estimated risk and survival functions are reported in each node. Only the ratios R_3 and R_4 are used in the final tree. In accordance with the MAD-log DTPO model, lower values of R_4 are associated with an increase risk of bankruptcy. Node 5 contains the riskier covariate pattern. It is formed by firm-years such that $R_4 < -0.447$ and $R_3 > 0.36945$. The fact that higher values of R_3 are associated with a higher risk is also apparent in the MAD-log model but its effect is not significant (p-value=0.124) there.

A single tree provides a convenient descriptive tool that may help to refine a parametric model. However, we are more interested here in comparing the two approaches (trees versus DTPO models). Hence, it is

Table 3: Four DTPO models for the bankruptcy data. The first value is the parameter estimate, the second one is the estimated standard error and the third one is the p-value.

	Original	Transformed ratios		
Parameter	Ratios	Truncated	log	MAD-log
α_1 (year 3)	-3.22	-3.90	-2.91	-3.67
	0.18	0.39	0.85	0.19
	<0.001	<0.001	<0.001	<0.001
α_2 (year 4)	-3.16	-3.98	-3.00	-3.82
	0.17	0.37	0.83	0.19
	<0.001	<0.001	<0.001	<0.001
α_3 (year 5)	-3.35	-4.32	-3.27	-4.16
	0.18	0.37	0.83	0.20
	<0.001	<0.001	<0.001	<0.001
α_4 (year 6)	-3.55	-4.49	-3.52	-4.39
	0.19	0.38	0.84	0.22
	<0.001	<0.001	<0.001	<0.001
α_5 (year 7)	-4.10	-4.88	-4.00	-4.83
	0.26	0.42	0.87	0.27
	<0.001	<0.001	<0.001	<0.001
α_6 (year 8)	-3.98	-4.85	-3.93	-4.78
	0.26	0.43	0.87	0.28
	<0.001	<0.001	<0.001	<0.001
R_1	-0.038	-0.022	-0.241	-0.116
	0.046	0.051	0.283	0.132
	0.409	0.663	0.395	0.378
R_2	-0.001	0.198	-0.034	0.138
	0.039	0.110	0.305	0.098
	0.977	0.071	0.910	0.156
R_3	**0.060**	**0.934**	-0.078	0.046
	0.025	0.390	0.583	0.124
	0.016	0.017	0.894	0.710
R_4	**-0.020**	**-3.676**	**2.611**	**-0.826**
	0.006	0.288	0.340	0.068
	0.001	<0.001	<0.001	<0.001
R_5	**-0.168**	**-0.062**	**-1.049**	**-0.749**
	0.046	0.016	0.161	0.133
	<0.001	<0.001	<0.001	<0.001
AIC	1499.0	1343.0	1419.8	**1331.8**
BIC	1573.0	1417.1	1493.9	**1405.8**

important to investigate the out-of-sample performance of the models. To do so, the model above along with forests of survival trees are now compared using ROC curves and a summary of the curves, the area under the ROC curve (AUC), via a cross-validation scheme. The 1143 firms were randomly divided into ten groups (10-fold cross-validation) in such a way that each group contains about 10% of the firms. But we did it in a stratified way such that each group contains also about 10% of the bankruptcies. Then the usual cross-validation paradigm was used for each model to be compared. More precisely, risk estimates were obtained for all observations in a group by fitting the model with the remaining groups. In the end and for each model, we have one out-of-sample estimated risk for each firm-year observation. These estimated risks are then used to compute the ROC curves and AUC.

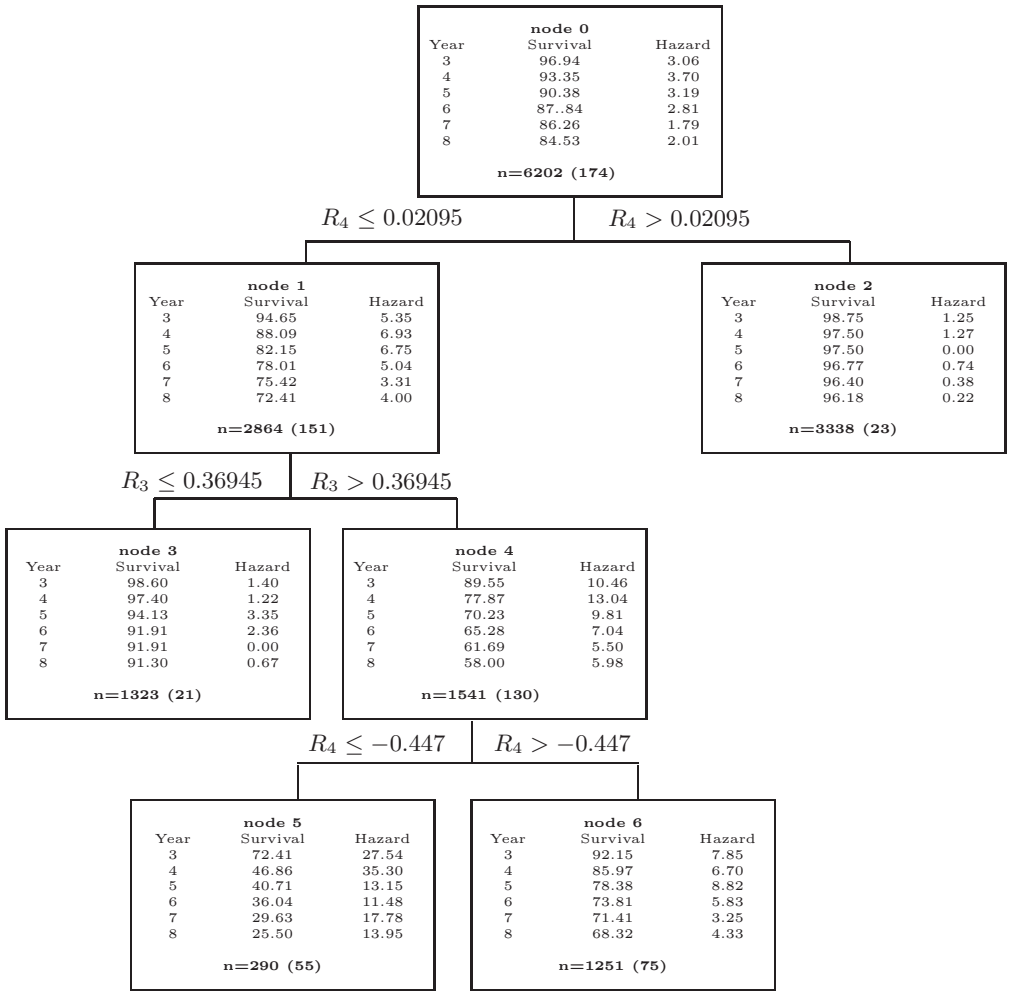

Figure 1: A single survival tree for the bankruptcy data. The estimated hazard and survival functions are reported in percent. "Year" is the number of years after the IPO. In each node, the total number of firm-year observations is given as "n=" and the number of bankruptcies is given between parentheses.

For the survival forest approach only the model when we select three out of five ratios in each node will be presented and discussed since it is the one that gave the best results. But straight bagging (choosing all ratios in each node) and the other survival forests provided very similar results. Each forest was built with 100 trees. Moreover, for the transformed ratios, only the MAD-log transformation will be presented as it is the best one in these out-of sample comparisons as it was also with the AIC and BIC criteria. Hence, we will be comparing four models: 1) the DTPO model with the original ratios, 2) the DTPO model with the MAD-log ratios, 3) the single tree and 4) the survival forest (with 100 trees) with three out of five ratios selected at random in each node.

Figure 2 presents the overall ROC curves for the four models. The corresponding AUC are reported in the upper part of Table 4. It is seen that the MAD-log DTPO model and the survival forest are better than the other two. The DTPO model with the original ratios seems to be the worse model and the single tree lies somewhere in between this one and the two top models. The AUC for each period are also reported in Table 4. The MAD-log and survival forest models are always the top two models in each period, the MAD-log being in first place for four out of six periods.

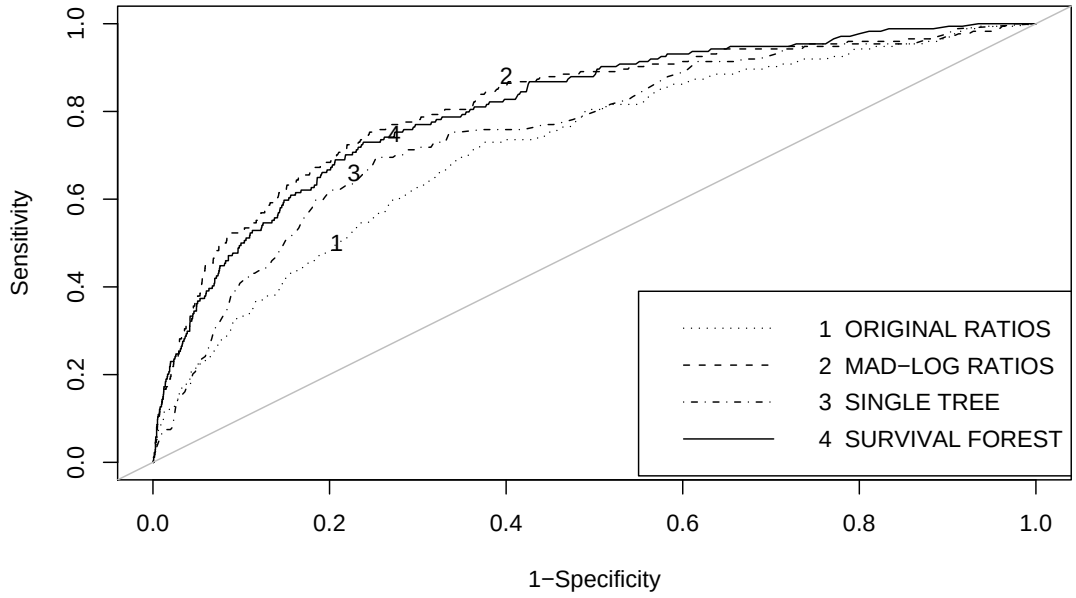

Figure 2: ROC curves for the out-of-sample risk estimates with the bankruptcy data. Four models are represented: DTPO models with the original and MAD-log ratios, a single tree and a survival forest with 100 trees.

Table 4: Area Under the ROC curves (AUC) for the out-of-sample risk estimates with the bankruptcy data.

Year after IPO	DTPO		Trees	
	Original	MAD-log	Single	Forest
All years combined	0.720	**0.814**	0.757	0.810
3	0.703	**0.802**	0.591	0.760
4	0.750	**0.834**	0.718	0.828
5	0.745	0.850	0.763	**0.855**
6	0.666	**0.742**	0.690	0.727
7	0.744	0.821	0.774	**0.844**
8	0.727	**0.804**	0.715	0.781

Overall, the MAD-log transformation provided a better model than the one using the original ratios. However, finding a good transformation is not a trivial task. We tried many transformations here and were fortunate to find what seems to be a reasonable one. At the same time, the performance of the survival forest is very close to the one of the MAD-log model. But the advantage of the survival forest approach lies in the fact that almost no intervention from the analyst is needed.

4 Concluding remarks

The motivating data for this work was bankruptcy data. Modeling bankruptcy data has a long history and the studies evolved from using single-period approaches to multiple-period approaches through survival analysis models. Discrete-time survival analysis methods are most often used because the status of each firm along with the usual covariates are yearly measures.

At the same time, survival trees became a widely accepted alternative to (semi) parametric models for the analysis of time-to-failure data. However, the methods were mainly developed under a continuous survival variable framework. It is only recently (Bou-Hamad et al., 2009) that a survival tree method specifically adapted for a discrete-time variable was proposed. However, this method could only incorporate time independent covariates. Hence, the method could not be applied to bankruptcy data studies that incorporate time-varying covariates such as annual financial and accounting based ratios. The purpose of this work was thus to generalize the Bou-Hamad et al. (2009) method to be able to use such time-varying covariates. One of benefits of the proposed method is that it allows both time-varying effects and time-varying covariates to be incorporated at the same time. Morever, since the splitting criterion has a closed-form, computation time is not an issue and we can easily build many trees to construct a forest of trees for instance.

Trees can be useful in a large variety of situations. A single tree can be an interesting descriptive tool in itself. Moreover, it can provide insights on the interactions among the covariates and help the analyst in the parametric model-building process. Sometimes a single tree can also be a good predictive tool. However, it is often the case that the combination of many trees will offer a better predictive performance than a single tree. Forest of trees (with bagging as a special case) are such methods that often provide very good out-of-sample predictive accuracies. Moreover, these methods are basically "off-the-shelf" since very little input from the analyst is needed. Discovering important interactions and/or finding appropriate covariate transformations is not a trivial task when using more classical parametric models and often involves a trial-and-error approach that needs many inputs from the analyst. Moreover, the variability involved with such ad-hoc model selection is rarely taken into account (because it is a difficult task) when we estimate the performance of a model. But the price to pay with methods like forest of trees is that the interpretation of the model is more difficult. If interpretation is of the foremost importance, than a model like a survival forest can at least serve as a benchmark to compare the performance of more interpretable models.

There are many possibilities for future work. For instance, improving the interpretability of forests of trees is still an ongoing research topic. In the context of this paper, an added difficulty comes from the fact that the effects of the covariates are time dependent. Another direction would be to develop a boosting approach adapted to discrete-time survival data with time-varying covariates using the tree method introduced in this paper. Finally, another possibility would be to compare the splitting criterion proposed in this paper to other ones based on restricted models like (6). Specifically the performance of different methods, including existing ones for continuous survival variables, could be investigated as the number of period increases in order to provide guidelines to practitioners.

References

Altman, E. I. (1968). Financial Ratios : Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance* **23**, 589–609.

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.

Beaver, W. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research* **5**, 71–111.

Bellovary, J. L., Giacomino, D. E. and Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education* **33**, 1–42.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., Mâsse, L., Vitaro, F. and Tremblay, R. (2009). Discrete-Time Survival Trees. *To appear in Canadian Journal of Statistics*.

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

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

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

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.

De Leonardis, D. and Rocci, R. (2008). Assessing the Default Risk by Means of a Discrete-time Survival Analysis Approach. *Applied Stochastic Models in Business and Industry* **24**, 291–306.

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

Doornik, J. A. (2002). *Object-Oriented Matrix Programming Using Ox*, 3rd edition. London: Timberlake Consultants Press and Oxford: www.doornik.com.

Fan, J., Su, X.-G., Levine, R. A., Nunn, M. A. and LeBlanc, M. (2006). Trees for Correlated Survival data by Goodness of Split, With Applications to Tooth Prognosis. *Journal of the American Statistical Association* **101**, 959–967.

Fanning, K. and Cogger, K. O. (1994). A Comparative Analysis of Artificial Neural Networks Using Financial Distress Prediction. *Intelligent Systems in Accounting, Finance and Management* **3**, 241–252.

Frydman, H. Altman, E. I. and Kao, D. (1985). Introducing Recursive Partitioning for Financial Classification : The Case of Financial Distress. *Journal of Finance* **40**, 269–291.

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

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

Hamza, M. and Larocque, D. (2005). An Empirical Comparison of Ensemble Methods Based on Classification Trees. *Journal of Statistical Computation and Simulation* **75**, 629–643.

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

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.

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.

Meyer, P. A. and Pifer, H. (1970). Prediction of Bank Failures. *Journal of Finance* **25**, 853–868.

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.

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.

Nam, C. W., Kim, T. S., Park, N. J. and Lee, H. K. (2008). Bankruptcy Prediction Using a Discrete-Time Duration Model Incorporating Temporal and Macroeconomic Dependencies. *Journal of Forecasting* **27**, 493–506.

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* **18**, 109–131.

R Development Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: www.R-project.org.

Ross, S. A., Westerfield, R. W.,Jordan, B. D. and Roberts, G. S. (2002) *Fundamentals of Corporate Finance*. Fourth Canadian Edition, McGraw-Hill Ryerson.

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.

Singer, J. D. and Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics* **18**, 155–195.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *Journal of Business* **74**, 101–124.

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

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

Zmijewski, M. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* **22**, 59–82.