Discrete-Time Survival Trees

Par

Imad Bou-Hamad

École des Hautes Études Commerciales

Thèse présentée à la faculté des études supérieures

en vue de l'obtention du grade de doctorat (PhD)

en administration,

option: Statistique Appliquée

Mai 2009

École des Hautes Études Commerciales

Cette thèse est intitulée :

Discrete-Time Survival Trees

Présentée Par Imad Bou-Hamad

a été évaluée par un jury composé des personnes suivantes :

Michèle Breton

.....

Président –rapporteur

Denis Larocque

.....

Directeur de recherche

Hatem Ben-Ameur

.....

Co-directeur de recherche

Marc Fredette

.....

Membre du jury

Ernest Monga

.....

Examinateur externe

Jean Roy

.....

Représentant du doyen de la FES

RÉSUMÉ

Les méthodes d'arbres sont souvent utilisées dans des études comportant des temps de survie avec censure. Les méthodes existantes ont été développées pour traiter des temps de survie mesurés sur une échelle continue avec des covariables ne variant pas dans le temps. Par contre, des temps de survie mesurés sur une échelle discrète et des covariables pouvant varier dans le temps sont souvent présents en pratique. Cette thèse est formée de trois articles.

La méthode de base est présentée dans le premier article. Il s'agit d'une méthode spécifiquement adaptée au cas de variable de survie à temps discret.

Dans le second article, nous étendons la méthode de base afin d'inclure des covariables variant dans le temps. Cette méthode permet aussi d'avoir des effets variant dans le temps. Nous présentons un exemple où nous étudions les facteurs reliés à la faillite à l'aide d'un échantillon de firmes américaines.

Finalement, le troisième article présente une revue des développements des méthodes pour données de survie avec censure.

Mots-clés : Arbres de survie; Données censurées à droite; Analyse de survie à temps discret; Forêts de survie; Covariables variant dans le temps; Effet variant dans le temps; Bagging; Données de faillite.

SUMMARY

Tree-based methods are frequently used in studies with censored survival time. The existing methods are tailor-made to deal with a survival time variable that is measured continuously and almost all of them cannot handle time-varying covariates. However, survival variables measured on a discrete scale and time-varying covariates are often encountered in practice. This thesis is composed of three articles.

In the first one, we propose our basic method which is a new tree construction method specifically adapted to discrete-time survival variables. The splitting procedure can be seen as an extension, to the case of right-censored data, of the entropy criterion for a categorical outcome.

In the second article, we extend the basic method to include time-varying covariates which are frequently encountered in practice. This method can accommodate simultaneously time-varying covariates and time-varying effects. We apply the new method to study the factors related to bankruptcy with a sample of United States firms that conducted an Initial Public Offerings between 1990 and 1999.

Finally, the third article presents a non-technical review of the developments of tree-based methods for the analysis of survival data with censoring.

Keywords: Survival trees; CART; Time--varying covariate; Right--censored data; Discrete-time survival analysis; Survival forests; Time-varying covariate; Time--varying effect; Bagging; Bankruptcy data;

TABLES DES MATIÈRES

LISTE DES TABLEAUX

PAPER 1

PAPER 2

LISTE DES FIGURES

PAPER 1

PAPER 2

REMERCIEMENTS

Je voudrais exprimer toute ma gratitude envers toutes les personnes qui m'ont encouragé et qui ont contribué de près ou de loin à la réalisation de cette thèse de doctorat.

Tout d'abord, je tiens à remercier chaleureusement mes co-directeurs de recherche Denis Larocque et Hatem Ben-Ameur pour leur précieuse aide, leurs commentaires judicieux et leur support financier. Je remercie particulièrement Denis, pour sa grande disponibilité, son implication, sa patience et ses discussions et conseils.

Mes remerciements s'adressent également aux membres du jury qui ont eu la gentillesse de lire et évaluer ma thèse.

Je n'oublie pas à remercier Mohamed Jabir pour son support informatique et ses explications pertinentes pour la base de données de COMPUSTAT.

Un remerciement particulier au docteur Antonio Ciampi pour son acceptation de lire et commenter une partie de mon travail.

Je désire aussi remercier les docteurs Michèle Breton et Georges Zaccour pour leur support moral tout au long de mon travail.

Les mots ne suffisent pas pour remercier les membres de ma famille, notamment ma mère et mon épouse et ainsi ma belle mère, pour leur amour, leur patience, leur encouragement et pour m'avoir soutenu dans les moments difficiles. Finalement, un grand bisou à mes petites chouettes Liyana et Rosalina.

Cette recherche à été supportée financièrement par le Centre de Recherche sur la E-Finance (CREF), le Conseil de Recherche en Sciences Naturelles et en Génie du Canada (CRSNG) ainsi que par le Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT)

INTRODUCTION GÉNÉRALE

La première application de méthodes d'arbres est apparue avec Morgan et Sonquist (1963) dans le contexte de la régression. Cependant, l'importante contribution théorique et appliquée de Breiman et al. (1984) a rendu ces méthodes très populaires tant pour l'exploration de données que pour le développement de modèles prévisionnels. Par l'entremise d'un algorithme de partitionnement récursif, un arbre partitionne l'espace des covariables afin de créer des classes homogènes par rapport à la variable réponse. À l'origine, les méthodes d'arbres ont été développées pour modéliser une variable réponse catégorielle univariée (arbre de classification) ou continue (arbre de régression). Cependant, elles ont été étendues à une large variété de situations dont les données longitudinales (Segal, 1992), les données multivariées (Zhang, 1998) et les données censurées (Gordon et Olshen, 1985). Cette dernière extension est connue sous le nom d'arbre de survie.

En général, un arbre de survie sert à regrouper les individus (ou unités statistiques) en classes homogènes par rapport au comportement de survie. Ce regroupement peut être utilisé afin d'identifier les facteurs de pronostique. En outre, un arbre de survie peut fournir des estimations des probabilités de survie et de la fonction de risque en fonction des covariables.

Les premières applications des arbres de survie sont apparues avec Gordon et Olshen (1985). Depuis ce temps, plusieurs méthodes ont été développées mais les travaux se sont concentrées sur le cas où la variable représentant le temps de survie est continue. Cependant, les variables de survie mesurées sur une échelle discrète sont souvent présentes en pratique. Par ailleurs, la quasi-totalité des méthodes d'arbres de survie traitent le cas de covariables qui sont fixes dans le temps. Seulement Bacchetti et Segal (1995) et Huang, Chen et Soong (1998) ont proposé des méthodes d'arbres de survie qui permettent l'incorporation de covariables qui varient dans le temps. Pourtant, on rencontre en pratique de nombreuses situations comportant des covariables

qui varient dans le temps. Une telle situation est l'étude des facteurs reliés aux faillites des entreprises.

De nombreuses méthodes statistiques ont été utilisées pour prévoir ou prévenir un stress financier. Les plus fameuses ont été l'analyse discriminante (Altman, 1968), la régression linéaire (Meyer et Pifer, 1970), la régression logistique (Ohlson, 1980), les modèles probit (Zmijweski, 1984), les arbres de décision (Frydman , Altman et Kao, 1985) et les réseaux de neurones (Fanning et Cogger, 1994). Cependant, ces études utilisent des modèles à une période dans le sens que la faillite ou non d'une entreprise est mesurée à un seul moment. Ils tentent alors de prévoir la faillite dans un horizon de temps fixe (allant de un à trois ans habituellement). Les études plus récentes ont utilisés des modèles d'analyse de survie afin de pouvoir inclure plusieurs périodes par entreprise.

Deux raisons principales ont motivé cette recherche : (1) le fait qu'il n'y a pas de méthodes d'arbre développées spécifiquement pour les temps de survie mesurés selon une échelle discrète lorsque le nombre de périodes est relativement petit et (2), la nécessité d'avoir une méthode permettant d'inclure des covariables qui varient dans le temps. Cette méthode pourra alors être une alternative intéressante aux méthodes utilisées dans la prévision de la faillite ou du stress financier des entreprises.

Cette thèse est formée de trois articles et est organisée comme suit. Dans le premier article, nous proposons une nouvelle méthodologie d'arbre de survie spécialement adaptée aux temps de survie mesurés selon une échelle discrète avec un petit nombre de périodes. Notre critère de séparation est basée sur l'approche de maximum de vraisemblance pour un modèle de survie paramétrique à temps discret. La taille de l'arbre est choisie selon un algorithme d'élagage combiné avec une validation croisée basée sur le bootstrap. Une investigation de la performance de la méthode proposée est effectuée via une simulation. De plus, une illustration pratique de cette méthode est

faite en analysant des données sur le tabagisme chez les adolescents, présentée dans l'étude de Mâsse et Tremblay (1997). Dans le deuxième article, nous proposons une méthode d'arbre de survie à temps discret qui admet des covariables qui varient dans le temps. Cette méthode généralise la méthode du premier article. Elle est appliquée à un échantillon d'entreprises américaines qui ont effectuée un premier appel public à l'épargne entre 1990 et 1999. Le troisième article présente une revue détaillée des méthodes d'arbres de survie proposées dans la littérature. Il se concentre sur les éléments fondamentaux tels les critères de séparation et les méthodes de sélection d'un arbre final. Il met aussi l'accent sur les développements récents tels les méthodes pour données de survie multivariées, l'utilisation de méthodes d'ensemble avec les arbres de survie ainsi que certains sujet précis comme les covariables et les effets variant dans le temps. Nous terminons la thèse par une conclusion générale.

Paper 1: Discrete-Time Survival Trees

Imad Bou-Hamad*, Denis Larocque*, Hatem Ben-Ameur*,
Louise C. Mâsse**, Frank Vitaro*** and Richard E. Tremblay***

* Department of Management Sciences
HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine,
Montréal, QC, Canada H3T 2A7

** Centre for Community Child Health Research
The University of British Columbia, L309-4480 Oak Street,
Vancouver, BC, Canada V6H 3V4

*** Research Unit of Children's Psycho-Social Maladjustment
University of Montréal, 3050, Édouard-Montpetit,
Montréal, QC, Canada H3T 1J7

This paper is published in Canadian Journal of Statistics (2009).

ABSTRACT

Tree-based methods are frequently used in studies with censored survival time. Their structure and ease of interpretability make them useful to identify prognostic factors and to predict conditional survival probabilities given an individual's covariates. The existing methods are tailor-made to deal with a survival time variable that is measured continuously. However, survival variables measured on a discrete scale are often encountered in practice. We propose a new tree construction method specifically adapted to such discrete-time survival variables. The splitting procedure can be seen as an extension, to the case of right-censored data, of the entropy criterion for a categorical outcome. The selection of the final tree is made through a pruning algorithm combined with a bootstrap correction. We also present a simple way of potentially improving the predictive performance of a single tree through bagging. A simulation study shows that single trees and bagged-trees perform well compared to a parametric model. A real data example investigating the usefulness of personality dimensions in predicting early onset of cigarette smoking is presented.

Keywords: Discrete-time survival analysis; Survival tree; Bagging; Maximum likelihood; Bootstrap.

1. INTRODUCTION

Tree-based methods have appeared with Morgan and Sonquist (1963) in the regression setting. However, the popularity of these methods has increased greatly with the important theoretical and practical contribution of Breiman, Friedman, Olshen and Stone (1984) coupled with the availability of the computing resources needed to apply them. Such recursive partitioning methods have been introduced as alternatives to classical parametric models like linear regression and discriminant analysis. Some advantages of tree-based methods are that they work on fewer assumptions compared to classical counterparts, they can automatically detect certain types of interactions, they can handle missing data and they produce sets of rules that can easily be understood and interpreted. Originally, tree-based methods were developed to model a univariate categorical response (classification tree) or a univariate continuous response (regression tree) but they have been extended to a wide variety of situations over the last two decades, e.g. for longitudinal data (Segal, 1992), for multivariate binary responses (Zhang, 1998) and for the case of interest in this paper, censored survival data, as discussed next.

A survival tree can serve to obtain a grouping of individuals, in terms of the values taken by their covariates, such that each group has a distinct survival behavior. Such a grouping can then be used to identify prognostic factors. Moreover, a survival tree can also be used to obtain survival probability estimates for new patients based on their covariate pattern.

The first wave of tree-based methods for survival data began with Gordon and Olshen (1985) who proposed to partition the data by using different measures of distance (e.g. L^p, Hellinger) between two survival curves. Ciampi, Thiffault, Nakache, and Asselain (1986), Segal (1988) and LeBlanc and Crowley (1993) used the two-

sample log-rank statistic to assess separation between two nodes. In their discussion, Therneau, Grambsch and Fleming (1990) proposed to use martingale residuals from a Cox model as inputs for a regression tree algorithm. Davis and Anderson (1989) considered an exponential log-likelihood loss function for node splitting while LeBlanc and Crowley (1992) used a full likelihood approach.

Many of these survival tree algorithms were compared in Radespiel-Tröger, Rabenstein, Schneider and Lausen (2003) and Radespiel-Tröger, Gefeller, Rabenstein and Hothorn (2006). More recently, Keles and Segal (2002) provided an analytic relationship between the logrank and the martingale residual sum-of-squares split functions. Jin, Lu, Stone and Black (2004) proposed a new splitting method by using a degree of separation index based on the variance of mean survival time. Su and Tsai (2005) proposed a hybrid model that combines a Cox's proportional hazards model and a tree. Combinations of several trees with bagging was proposed in Hothorn, Lausen, Benner and Radespiel-Tröger (2004) and with random forest and boosting in Hothorn, Bühlmann, Dudoit, Molinaro and Van Der Laan (2006) and Ishwaran, Kogalur, Blackstone and Lauer (2008). For the case of multivariate survival data, Su and Fan (2004) and Gao, Manatunga and Chen (2004) proposed methods based on frailty models and Molinaro, Dudoit and Van Der Laan (2004) are providing a unified methodology for right-censored data that includes the multivariate case. Finally, Fan, Su, Levine, Nunn and LeBlanc (2006) extended the LeBlanc and Crowley (1993) method to correlated survival data.

The previous studies about survival trees were mainly developed to deal with a continuous survival time variable. However, discrete survival time variables occur often in practice. When the number of observed times is large and not too many ties are present, then treating them as continuous would be a reasonable approach. However,

when the number of observed times is small, then some specifically adapted methods are needed. The discrete-time proportional odds model proposed in Cox (1972) was made available to a large field of practitioners by Singer and Willett (1993). This popular model possesses many advantages: i) the hazards can be interpreted as conditional probabilities and not just as rates, ii) its basic assumption about proportional odds can easily be relaxed, iii) it can be fitted using any logistic regression software and, iv) it can easily accommodate time-varying covariates. For instance, Mâsse and Tremblay (1997) used this model to relate personality dimensions measured on boys at age 6 to the onset of cigarette smoking, alcohol abuse and other drug use during adolescence. This study will be revisited in this paper.

The fact that no tree-based methods were developed specifically to model a censored discrete-time survival variable when the number of observed time is small was the main motivation for this work. In this paper, we are (1) proposing a complete methodology to handle such situations with a tree-based approach based on maximum likelihood, (2) investigating the performance of the method through a simulation study, and (3) illustrating the use of the method by reanalyzing the data from Mâsse and Tremblay (1997). The paper is organized as follows. The tree construction algorithm is described in details in Section 2. More specifically, that section reviews the basic assumptions about the data, describes the splitting rules used for tree-building, explains how to select a final tree with a pruning and selection procedure and describes how to aggregate many trees through bagging. Section 3 presents the results from the simulation study that evaluates the performance of the proposed method. The real data example is presented in Section 4 followed by some concluding remarks.

2. TREE CONSTRUCTION

2.1. Data description and notation

Data is available for N individuals and the independent vectors $(\tau_i, \delta_i, \mathbf{x}_i)$, for $i = 1, \ldots, N$, are observed. The observed discrete survival time for the i^{th} individual, τ_i, is assumed to be a positive integer without loss of generality. The censoring indicator δ_i is defined as usual by $\delta_i = 0$ if the observed time of individual i is right-censored and $\delta_i = 1$ if the true time-to-event is observed. The vector of covariates for individual i is $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. These covariates can be a mix of continuous and categorical variables as both types of variables are allowed in the tree construction. As it is usually done, we assume that the true censoring time and true time-to-event are independent given the covariates. However, we are not making any type of proportional odds assumptions as needed in the basic discrete-time proportional odds model.

Let $U \in \{1, 2, \ldots\}$ be the real time-to-event for an individual chosen at random in the population under study. For $j = 1, 2, \ldots$, we denote the discrete-time hazards by $h(j)$, the survival probabilities by $S(j)$ and the probabilities of events by $\pi(j)$. Namely,

$$h(j) = P(U = j | U \geq j), \quad S(j) = P(U > j) \quad \text{and} \quad \pi(j) = P(U = j). \qquad (1)$$

For a given individual, these quantities depend on the covariate vector but we use this simplified notation since no confusion is possible. Assume that K is the maximum observed time for a given data set. A well-known model to analyze discrete-time survival data is the discrete-time proportional odds (DTPO) model (Singer and Willett, 1993). This model relates the hazards to the covariates as follows:

$$\log\left(\frac{h(j)}{1-h(j)}\right) = \alpha_1 D_1(j) + \alpha_2 D_2(j) + \cdots + \alpha_K D_K(j) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \quad (2)$$

for $j = 1, \ldots, K$, where the $D_k(j)$'s are indicator variables indexing the time periods that are defined by $D_k(j) = 1$ if $k = j$ and 0 otherwise. The intercept parameters $\alpha_1, \ldots, \alpha_K$ define the baseline of hazard in each time period and the β coefficients describe the effects of covariates on the baseline hazard function. One nice feature of this model is that the parameters can easily be estimated by maximum likelihood by using any logistic regression software with an appropriately modified data set. See Singer and Willett (1993) for a description of the model and its extensions and how to use it in practice. Note that the basic model assumes proportional odds but this assumption can be relaxed by introducing interactions terms between the covariates and the indicator variables.

2.2. Splitting rule

Tree-based methods partition the covariates space by splitting it recursively with rules based on covariates. Two components are needed in a tree construction algorithm: i) a splitting rule and ii) a method to select one tree from the sequence of trees generated by the splitting rule. This subsection describes the first component while the second component will be the subject of the following subsection. It is assumed that the reader is familiar with the basic CART method and its associated terminology (Breiman *et al.*, 1984).

As for the basic CART method, only binary splits on one covariate are considered. For a continuous covariate x, the possible splits take the form $x \leq c$ where c is a specified cutpoint. For a categorical covariate x, the possible splits take the form

$x \in \{c_1, \ldots, c_k\}$ where $\{c_1, \ldots, c_k\}$ is a subset of possible values of x. At a given node, all possible splits on all covariates are considered and a "best" split is chosen according to a criterion.

By adding a subscript to identify the individuals, we define $h_i(j)$, $S_i(j)$ and $\pi_i(j)$ as in (1), for $i = 1, \ldots N$. In this paper, we propose a splitting rule based on maximum likelihood in the spirit of Su, Wang and Fan (2004). Basically, the chosen split will be the one that maximizes the likelihood of a particular two-node model as described next.

Let t be a node of the tree. Let n_t be the number of observations and $J(t)$ the set of indices of the observations that are in node t. The likelihood at node t can then be written as

$$L(t) = \prod_{i \in J(t)} \pi_i^{\delta_i}(\tau_i) S_i^{1-\delta_i}(\tau_i). \tag{3}$$

Without any covariate, which amounts to using the intercept only model

$$\log\left(\frac{h(j)}{1-h(j)}\right) = \alpha_1 D_1(j) + \alpha_2 D_2(j) + \cdots + \alpha_K D_K(j), \tag{4}$$

the log-likelihood evaluated at the maximum likelihood estimates (MLEs), i.e. the observed log-likelihood, is given by

$$ll(t) = \sum_{j=1}^{K} \{n_{t1}(j) \ln(\hat{\pi}(j)) + n_{t0}(j) \ln(\hat{S}(j))\} \tag{5}$$

where $n_{t0}(j)$ and $n_{t1}(j)$ are the number of individuals in node t for which $\tau_i = j$ and $\delta_i = 0$ and 1 respectively, and where $\hat{\pi}(j)$ and $\hat{S}(j)$ are the MLEs of $\pi(j)$ and $S(j)$. See Section 10.2 of Singer and Willett (2003). They are also provided in the Appendix A1 for the sake of completeness.

Note that when the node contains only true times to event (i.e. no censored

observations), then $n_{t0}(j) = 0$, $\hat{\pi}(j) = n_{t1}(j)/n_t$ and (5) reduces to

$$n_t \sum_{j=1}^{K} \hat{\pi}(j) \ln(\hat{\pi}(j)),$$

which is the well-known entropy criterion used for a categorical response variable without censoring. See Section 9.2.3 of Hastie, Tibshirani and Friedman (2001).

Now we wish to split the node t into two nodes. Let I denote the indicator function. For a candidate split, let C be the binary variable defining the split. For example, $C = I(x \leq c)$ would be a candidate split on a continuous covariate while $C = I(x \in \{c_1, \ldots, c_k\})$ would be candidate split on a categorical covariate. Following the split, the node t would be divided into a left node t^L containing the subset of node t for which $C = 1$ and a right node t^R containing the subset of node t for which $C = 0$. One possibility would be to model the two nodes as

$$\log\left(\frac{h(j)}{1-h(j)}\right) = \alpha_1 D_1(j) + \cdots + \alpha_K D_K(j) + \beta C, \tag{6}$$

to fit it and then use the log-likelihood computed at the MLEs as a criterion. But this has two features that are undesirable for a tree model. Firstly, computing the MLEs of this model require the use of numerical methods as standard logistic regression. This could be computationally demanding for large data sets with a large number of covariates since the MLEs would have to be obtained at many candidate splits. Secondly, this model imposes proportional odds with respect to C which is an unappealing assumption for a "nonparametric" model like a tree. A simple way to get rid of these two features at the same time is to add interaction terms between C and the

indicator variables thus modeling the two nodes as

$$\log\left(\frac{h(j)}{1-h(j)}\right) = \alpha_1 D_1(j) + ... + \alpha_K D_K(j) + \beta_1 C D_1(j) + \cdots + \beta_K C D_K(j). \quad (7)$$

By doing so, it is straightforward to see that the contributions of the left and right nodes observations to the likelihood function are completely separated. Thus, fitting model (7) is equivalent to fitting an intercept model (4) separately in each node. This lifts the proportional odds assumption, allowing the variable C to have a time-varying effect. Moreover, it speeds up computations since numerical methods become unnecessary. The total observed log-likelihood of this two-node model with the split C is then

$$ll(t^L) + ll(t^R) =$$

$$\sum_{j=1}^{K}(n_{t1}^L(j)\ln(\hat{\pi}^L(j)) + n_{t0}^L(j)\ln(\hat{S}^L(j))) + \sum_{j=1}^{K}(n_{t1}^R(j)\ln(\hat{\pi}^R(j)) + n_{t0}^R(j)\ln(\hat{S}^R(j))),$$

where $n_{t0}^L(j)$, $n_{t1}^L(j)$, $\hat{\pi}^L(j)$ and $\hat{S}^L(j)$ are defined as for (5) but with only the observations from the left node t^L and are defined similarly for the right node t^R. The chosen best split is then the one for which $ll(t^L) + ll(t^R)$ is maximum. Equivalently, the best split is the one with the maximum value of

$$g(t) = -2(ll(t) - (ll(t^L) + ll(t^R)) \quad (8)$$

which is the likelihood ratio statistic for comparing the single node model (parent node) to the two-node model (children nodes). We will refer to $g(t)$ as the splitting statistic for node t from now on.

One drawback of using (7) as the two-node model is that the number of parameters is $2K$ as opposed to $K+1$ with the main effect two-node model (6). Since $K \leq 4$ in the

cases considered in this paper, only the splitting rule based on (7) was implemented and investigated. But the more parsimonious model could be useful for larger values of K despite its two undesirable features. However, when K is large, another possibility is to simply treat the observed times as continuous and use one of the available tree methods for such data. Hence, our method is really aimed at the case where K is small. Then using (7) is appealing since, once again, it does not impose proportional odds and it can be computed quickly. Hence, the proposed method lets the data speak without imposing unnecessary assumptions right from the start.

2.3. Pruning and selection of the final tree

The growth of a tree can in principle be continued as long as there are enough observations in a node to allow it to be splitted further. In practice, fully grown trees tend to overfit the data and this is why some sort of stopping criterion is needed. The selection of a single tree is a difficult problem because of its discrete nature and most methods will exhibit high variability with finite samples. Note that when the tree method is used as an exploratory tool, then choosing a single tree is not necessary.

However, if a single tree must be chosen, we are proposing a pruning approach, as in CART, in which a large tree is grown, then pruned back to produce a sequence of nested trees from which one member is selected as the final tree. More specifically, we are proposing to use the pruning and selection method of LeBlanc and Crowley (1993) and Fan *et al.* (2006) which we now describe the basic ideas. The technical aspects of the method is left for the Appendix. For a given tree A, let $W(A)$ denote the set of interior nodes (i.e., all nodes except the terminal nodes). The sum of all

splitting statistics for the tree A is

$$G(A) = \sum_{t \in W(A)} g(t). \tag{9}$$

The performance of A will be evaluated by the split-complexity measure

$$G_\alpha(A) = G(A) - \alpha |W(A)| \tag{10}$$

where $|W(A)|$ is number of interior nodes of A and α is a nonnegative penalty term. For a given value of α, the larger $G_\alpha(A)$ is, the better A is. We define $G(A_M)$ to be 0 where A_M is the root-only tree. When $\alpha = 0$, the largest tree is the best according to (10) and, when $\alpha \to \infty$, the root-only tree would eventually be the one maximizing (10). Starting from a large tree A_0, it is possible to obtain a sequence of nested trees $\{A_0, A_1, \ldots, A_M\}$ and a corresponding sequence of α values $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_M < \infty$ such that A_m is the tree maximizing (10) for any α in the interval $[\alpha_m, \alpha_{m+1})$. Details on how to obtain the sequence of trees and α values are given in Appendix A2.

In principle, the final tree is the one, among $\{A_0, A_1, \ldots, A_M\}$, that maximizes (10). But for this, a value of α must be selected. In LeBlanc and Crowley (1993) and Fan *et al.* (2006), the split statistic at a given node has asymptotically a χ^2_1 distribution. They suggest to use 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 (the 0.95 quantile is in fact 3.84). From that point of view, $\alpha = 2$ corresponds to the 0.84 quantile. In our case, the split statistic $g(t)$ has a χ^2_K distribution asymptotically since the two-node model possesses K more parameters than the one node model. Hence, we suggest

selecting the penalty term based on an appropriate χ^2_K quantile. Taking a quantile in the interval $[0.85, 0.95]$ would be the equivalent of the suggestion made by LeBlanc and Crowley (1993). In the simulation study of the next section, we used the 0.90 quantile.

Once the penalty term is chosen, there is still one problem associated with the use of $G(A)$. The initial tree and the sequence of nested trees are obtained from the same sample. Using the same sample again to compute the $G(A)$'s is likely to produce too large (optimistic) values since the tree building algorithm seeks to maximize them at each node. Corrected and "honest" values of the $G(A)$'s are thus needed.

For large sample sizes, it is possible to split the original sample into a training and a test samples. The training sample is used to grow an initial large tree and prune it to obtain the sequence of nested trees. Then the split statistics are recomputed with the test sample to produce corrected values $G^{(c)}(A)$. To maximize the use of the data when the sample size is small or moderate, a bootstrap correction method is preferred. This is also the approach proposed in LeBlanc and Crowley (1993) and Fan *et al.* (2006). This method is explained in Appendix A2 and it also produces (bootstrap) corrected values $G^{(c)}(A)$. Once corrected values are calculated, whether from split samples or from bootstrap resampling, they are then used in conjunction with the chosen penalty term to select a final tree with the split complexity measure (10). Namely, the chosen tree is the one, among $\{A_0, A_1, \ldots, A_M\}$, which maximizes

$$G^{(c)}(A) - \alpha|W(A)|. \tag{11}$$

Once the final tree is selected, final estimates of the quantities of interest (hazard, survival probabilities and so on) can be obtained from all data points.

2.4. Bagging

Bagging is an ensemble method that works by averaging together the outputs from many trees built with bootstrapped data sets obtained from the original data. The method was introduced by Breiman (1996) and was successfully applied to many problems. Its main benefit is usually to reduce the variance associated with a single tree but the price to pay is that bagged trees are not as easily interpreted as a single tree. Hothorn *et al.* (2004) used bagging with survival data. To predict a new observation, their algorithm aggregates all training observations falling into the same terminal nodes as the one to be predicted and then computes the Kaplan-Meier curve of these observations.

In this paper we use bagging in a slightly different way compared to Hothorn *et al.* (2004) in the sense that the estimated probabilities of event are computed for each tree, then averaged over the trees and finally estimated survival probabilities are computed from these averaged probabilities. More specifically, our bagging algorithm works as follows:

1. Draw B bootstrap samples from the original data set.

2. For each bootstrap data set, grow a tree with the splitting rule described in Section 2.2. No pruning is performed. The splitting is stopped when a minimum node size is reached. In general, this size should depend on the number of censored and uncensored observations in the training sample. For simplicity, in the applications of this paper, a node is not splitted further and becomes a terminal node when it contains less than 40 observations.

To obtain estimated survival probabilities for a new observation:

1. Throw down the observation in each of the B trees. Let $\hat{\pi}^b(j)$ denote the estimate of $\pi(j)$, $j = 1, \ldots, K$, obtained from the b^{th} tree, $b = 1, \ldots, B$.

2. Let $\hat{\pi}(j) = \frac{1}{B}\sum_{b=1}^{B} \hat{\pi}^b(j)$ denote the final bagged estimate of $\pi(j)$.

3. The bagged estimate of the survival probabilities are then calculated recursively as $\hat{S}(1) = 1 - \hat{\pi}(1)$, and $\hat{S}(j) = \hat{S}(j-1) - \hat{\pi}(j)$ for $j = 2, \ldots, K$.

A more elaborate version of bagging called random forests (Breiman, 2001, Hothorn *et al.*, 2006 and Ishwaran *et al.*, 2008), in which only a random subset of covariates is selected at each node, could also easily be implemented. This method has been shown empirically to often be superior to bagging (Hamza and Larocque, 2005) and it also speeds up computations by reducing the number of splits to be evaluated. But straight bagging was used in this paper since the situations considered involve only a small number of covariates.

3. SIMULATION STUDY

3.1. Description of the study

The relative performance of three models is investigated in this section. These models are i) a single tree selected with the bootstrap correction method, ii) bagging with 100 trees and iii) the DTPO model defined in (2) using all covariates. The tree building algorithm is implemented in Ox (Doornik, 2002), while maximum likelihood estimation of the DTPO model is implemented in R (R Development Core Team, 2007).

Four different data generating processes (DGP) were used. The first one is a proportional odds model of the form (2) and the other three are tree-based models.

For each DGP, the observed time is either 1, 2 , 3 or 4, i.e. $K = 4$. In all cases, the covariates are generated independently of each other as uniform random variables in the interval $[0, 10]$. To add noise, three additional unrelated covariates were also generated for each DGP. For instance, if a DGP is defined with 2 covariates, then a total of five covariates were given to the models.

DGP 1 is defined through model (2) and is given by

$$\log\left(\frac{h(j)}{1-h(j)}\right) = 3D_1(j) + D_2(j) - D_4(j) + X_1 - 2X_2 \quad \text{for} \quad j = 1, 2, 3, 4.$$

For this DGP, we used a real censoring time variable with vector of probabilities $(.05, .1, .15, .7)$.

DGPs 2, 3 and 4 are tree-based models whose shapes and real survival probabilities and hazards in each terminal node are provided in figures 1, 2 and 3. The vector of probabilities for the real censoring time is the same in each terminal node for those three DGP and is given by $(.04, .08, .12, .76)$.

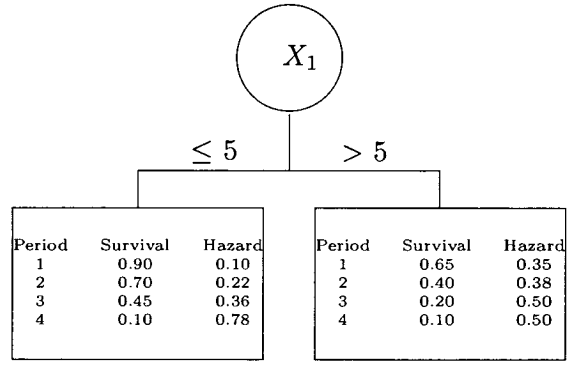

Figure 1: DGP 2 for the simulation study

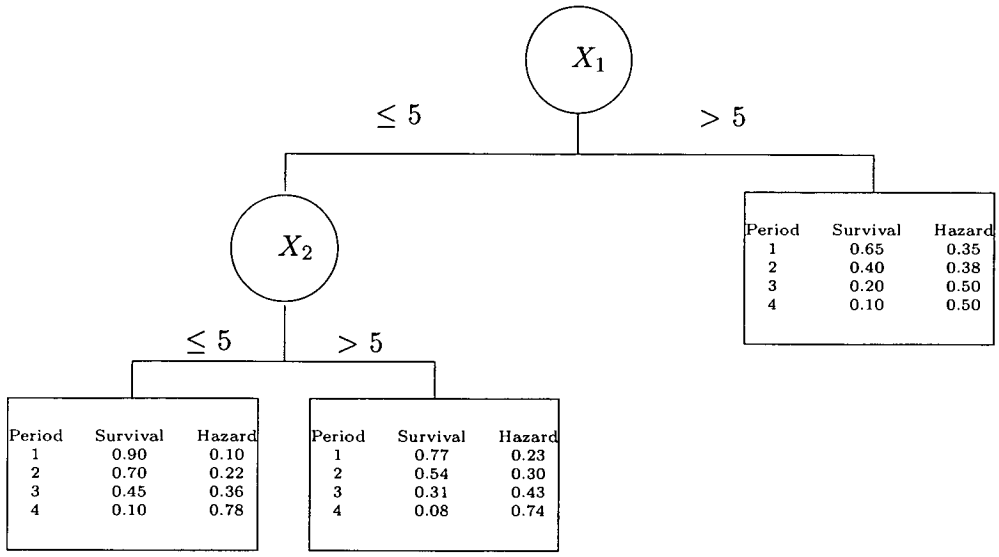

Figure 2: DGP 3 for the simulation study

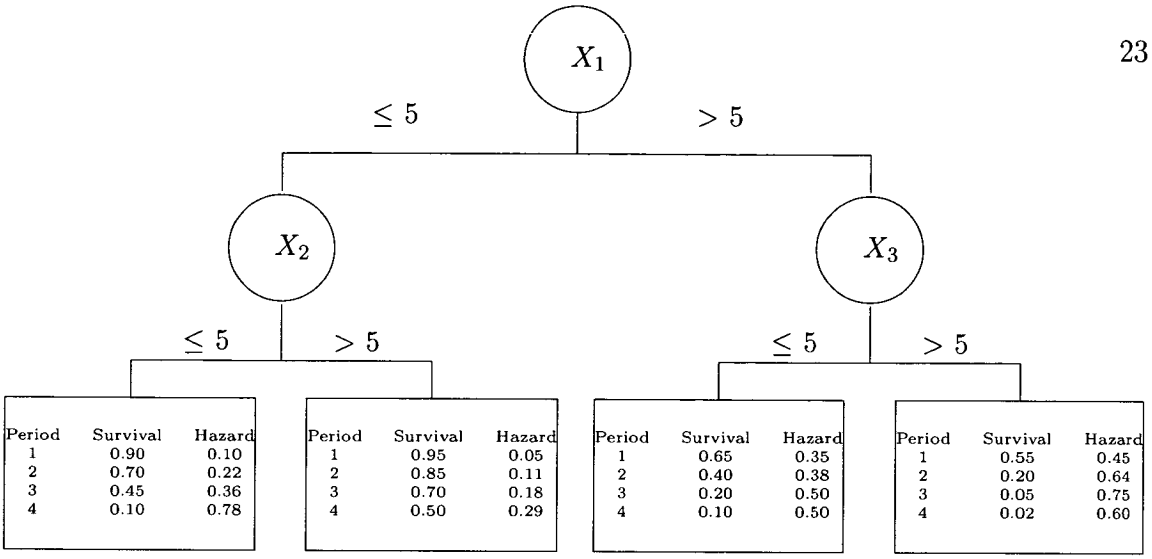

Figure 3: DGP 4 for the simulation study

With these parameters, the proportions of censored observations is about 19% 18%, 25% and 26% for DGP 1, 2, 3 and 4 respectively.

For each DGP, 1000 samples of size $n = 600$ were generated and the three models were fitted to each sample. The performance of any given model was evaluated through three criteria. But since the conclusions reached by following any of these criteria were very similar, we will only present the results for one of them, the mean absolute error (MAE) defined by

$$\frac{1}{4}\sum_{j=1}^{4}|\hat{S}(j) - S(j)| \tag{12}$$

where $S(j)$ is the real survival probability for time j and $\hat{S}(j)$ is an estimation of $S(j)$ from a fitted model. For each DGP and each generated sample, an estimate of MAE was obtained with an independent test set (not used in model fitting) of size 5000. These estimates were then averaged over the 1000 samples to produce the final performance criterion. The other two criteria that produced similar conclusions are $1/4\sum_{j=1}^{4}(\hat{S}(j) - S(j))^2$ and $1/4\sum_{j=1}^{4}(\hat{S}(j) - S(j))^2\pi(j)$.

For the single tree, the chosen penalty term in (11) is 7.78, the 0.90 quantile of the

χ^2_4 distribution. The bootstrap correction was performed with 30 samples each time.

3.2. Results

The results of the simulation are summarized in Table 1 and Figure 4. They present the boxplots of the MAE for the 1000 simulated samples for each DGP and model and some summary statistics.

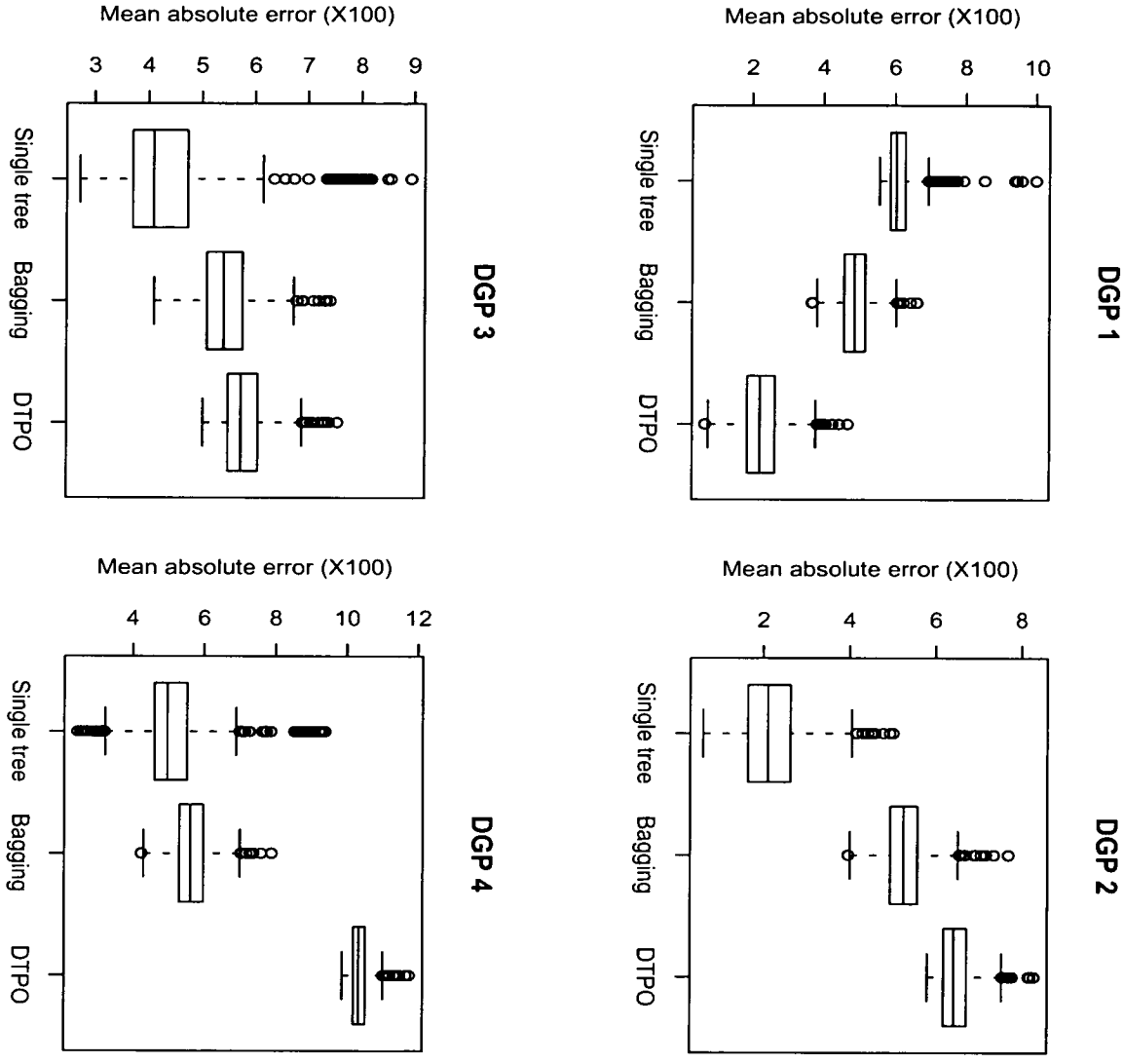

Figure 4: Results from the simulation study

Each graph provides the distribution of the estimated mean absolute error (MAE) for the 1000 simulated data sets for each data generating process (DGP). The three competing methods are a single tree selected with the pruning and bootstrap method, bagging with 100 trees and a discrete-time proportional odds (DTPO) model using all covariates.

Table 1: Summary statistics for the mean absolute error (MAE) for the simulation study (1000 runs)
All entries are multiplied by 100.

DGP	Model	Average	Std Error
1	Single tree	6.15	0.015
	Bagging	4.89	0.014
	DTPO	2.25	0.019
2	Single tree	2.17	0.023
	Bagging	5.30	0.016
	DTPO	6.50	0.013
3	Single tree	4.58	0.044
	Bagging	5.45	0.016
	DTPO	5.80	0.014
4	Single tree	5.37	0.042
	Bagging	5.67	0.016
	DTPO	10.37	0.009

As expected, the DTPO model is the one with smallest MAE for DGP 1. It has an average MAE of 0.0225. Bagging follows in second place with an average MAE of 0.0489. The single tree achieves the best performance for the three other DGPs which is not surprising since they are tree-based DGPs. The DTPO model is doing consistently a poor job for the more complex tree DGP (DGP 4). Bagging always comes in second place in all DGPs. We should not be disappointed by the performance of bagging in this simulation study because three of the four DGPs are tree-based models. It is then natural that a single tree does better than bagging. But when the DGP is not a tree, like for DGP 1, then bagging does better than a single tree. Moreover, when we move from DGP 2 to DGP 4, that is, from a simple tree DGP to a more complex tree DGP, then bagging gets closer and closer to the single tree.

4. DATA EXAMPLE AND CONCLUDING REMARKS

4.1. Onset of cigarette smoking

In this section, we provide an example illustrating the use of the proposed tree-based method. The data come from a longitudinal study on white boys from low socioeconomic French-speaking schools in the city of Montréal in the province of Québec in Canada. These data were analyzed in Mâsse and Tremblay (1997). Only a general description of the study is provided here and we refer the reader to the original paper for more details. One purpose of the study was to investigate the relation between three personality dimensions, novelty-seeking, harm-avoidance and reward dependence, and the onset of cigarette smoking, alcohol abuse and other drug use in boys. The possible values for the personality dimension variables are 1 (low), 2 (medium) and 3 (high). According to Cloninger's theory (Cloninger, 1987), higher scores of novelty-seeking, lower scores of harm-avoidance and lower scores of reward dependence are hypothesized to be predictors of early onset of alcohol-seeking behavior but other authors have shown that they are also related to other types of substances used and abused in adolescents. For our example, only the onset of cigarette smoking is treated. The three Cloninger's personality dimensions scores are available when the boys were 6 years old. The outcome of interest, whether the boys had smoked cigarettes, for the first time, in the past 12 months were assessed at ages 13, 14 and 15. Mâsse and Tremblay (1997) had 656 subjects for their analysis but our sample contains 740 boys since information on new subjects were made available after their initial study was concluded. The data are treated as discrete-time survival data with three time periods (13, 14 and 15 years old). By age 15 years, 40.3% (298 out of 740) of the boys reported having smoked a cigarette and the 59.7% who did not are considered censored

Table 2: Results for the onset of cigarette smoking example
Three discrete-time proportional odds (DTPO) models are presented. For each model, the first value is the estimate of the parameter and the value between parentheses is its estimated standard error. The first model contains all covariates. The second model is the one selected by the AIC criterion. The third model is the one selected by the BIC criterion. A * indicates a p-value<0.01.

	Model		
Parameter	Full model	AIC model	BIC model
α_1 (age 13)	-1.76(0.27*)	-1.83(0.21*)	-2.16(0.18*)
α_2 (age 14)	-2.23(0.28*)	-2.31(0.23*)	-2.64(0.20*)
α_3 (age 15)	-1.41(0.27*)	-1.48(0.21*)	-1.82(0.18*)
Novelty-seeking	0.26(0.07*)	0.26(0.07*)	0.24(0.07*)
Harm avoidance	-0.19(0.07*)	-0.19(0.07*)	
Reward dependence	-0.03(0.07)		
AIC	1625.6	**1623.8**	1629.0
BIC	1659.0	1651.6	**1651.2**

observations.

Mâsse and Tremblay (1997) analyzed the data through a DTPO model (2). Even though the covariates are trichotomous, they were modeled as linear effects by Mâsse and Tremblay (1997) and we are also doing so to reproduce their analysis. Three DTPO models were fitted to the data and Table 2 presents the results. The first model in Table 2 is the equivalent of the fifth model reported in Table 1 of Mâsse and Tremblay (1997). Since we are doing the analysis with a different sample, our coefficients differ slightly from the ones of the original study but the signs and significance (at a 5% level) of the parameters are identical. Only novelty-seeking and harm avoidance are significantly related to cigarette smoking and their effects are in accordance with Cloninger's theory. The second and third models are the ones selected by the AIC and BIC criteria respectively. We see that only novelty-seeking is retained by the BIC criterion.

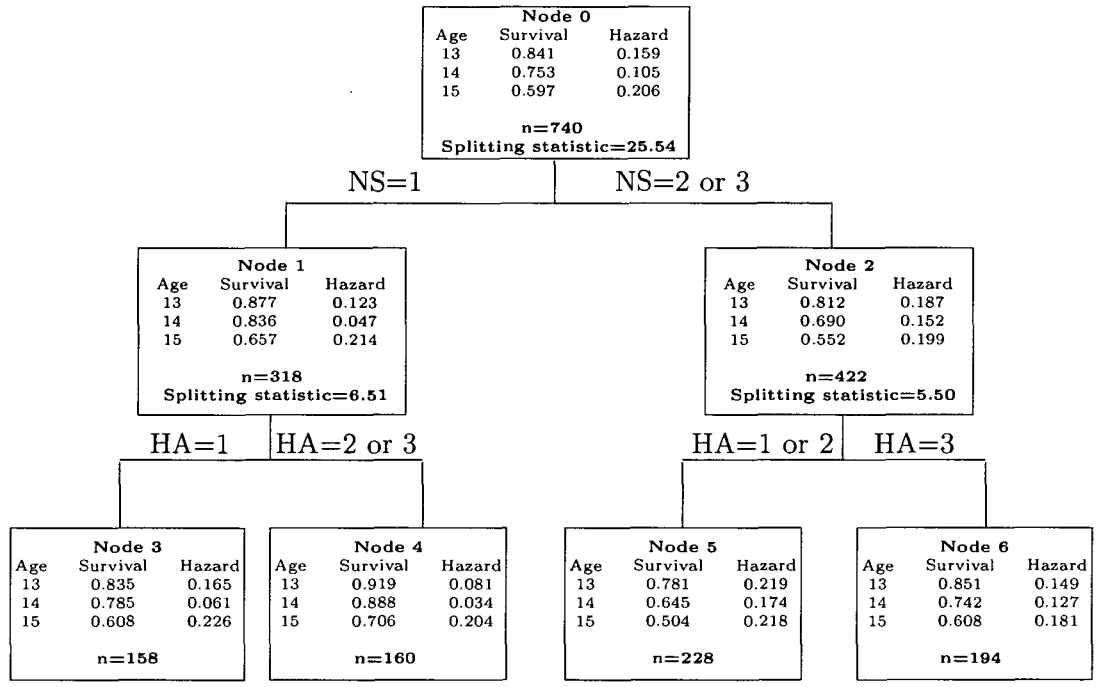

Figure 5: A discrete-time survival tree for the onset of cigarette smoking example

NS stands for Novelty-seeking and HA for Harm-avoidance. The tree with the two terminal nodes 1 and 2 (the stump) is the one retained by the pruning and bootstrap selection method.

We then built a tree using the method proposed in this paper. With a penalty term of 6.25, the 0.90 quantile of the χ^2_3 distribution, and 100 bootstrap samples, the simple stump (the tree with two terminal nodes) was selected as the best tree. But for illustration purpose, a larger tree with four terminal nodes in presented in Figure 5. The best tree does a single split on the covariate novelty-seeking. Boys with higher values (2 or 3) of novelty-seeking are more at risk of onset of cigarette smoking than those with the value 1, in accordance with Cloninger's theory. Even though the two other splits are not retained for the final tree, it is interesting to see that they are made with the covariate harm avoidance in accordance with the theory. Indeed, lower values of harm avoidance increase the risk in both splits.

Unlike what we did for the simulation study, we can not directly use the MAE (12) to compare models since the true S is unknown. Instead we used an empirical version of the MAE in conjunction with V-fold cross-validation. For the v^{th} cross-validation iteration ($v = 1, \ldots, V$), let $\hat{S}_i^v(j)$ be the estimated value of $S(j)$, obtained from the training sample, for the i^{th} observation in the test sample. The cross-validated empirical MAE is then

$$\text{MAE}_{cv} = \frac{1}{KV} \sum_{v=1}^{V} \frac{1}{m_v} \sum_{i \in T_v} \sum_{j=1}^{K} |\hat{S}_i^v(j) - I(\tau_i^v > j)|$$

where T_v is the test sample for iteration v, m_v is its size and τ_i^v is its i^{th} observation. It is important to say that each model was rebuilt entirely in each iteration. This means that the whole tree construction (building, pruning and selection via bootstrapping) was performed in each iteration using the training sample. By using 10-fold cross validation, it turns out that the cross-validated empirical MAE is only slightly lower ($\text{MAE}_{cv} = 0.367$) for the full DTPO model (the first one in Table 2) compared to a tree model ($\text{MAE}_{cv} = 0.369$). We also assessed the performance of bagging with 100

trees who did slightly better than the two other models ($MAE_{cv} = 0.366$).

To investigate the stability of the selected tree model, we tried different values for the penalty term. It turns out that the model is very stable since the stump with novelty-seeking would also be selected for all α values in the interval $[0.04, 20.2]$. This and the fact that only novelty-seeking is retained by the BIC criterion in the DTPO model seem to indicate that this covariate is the main one associated with the risk of onset of cigarette smoking in this population.

4.2. Concluding remarks

In conclusion, a new and complete methodology to build a tree for discrete-time survival data has been described in this paper. The splitting criterion of the proposed method reduces to the entropy criterion when there is no censoring. Hence, the new method can be seen as an extension of the classical classification tree method to the case of right-censored data. The simulation study has shown that the new method works well and its use was illustrated with a real data example. This methodology is mostly aimed at the case where the number of observed times is small since they could be treated as continuous otherwise. The method does not impose the proportional odds assumption frequently encountered in parametric and semi-parametric models. This has two benefits: i) time-varying effects for splitting variables are automatically incorporated and ii) the computation time is greatly reduced since the evaluation of the splitting criterion is based on a closed form expression.

A referee pointed out that using the two-node proportional odds (PO) model (6) is also possible in practice especially since fast score tests are available. The referee added that using such a model may be useful for power reasons since it contains less parameters. Thus, if the PO assumption holds, more accurate estimators could

be obtained especially lower in the tree where less observations are available. These comments are pertinent and interesting and certainly could justify future work. We think that not having to assume PO blindly is one of the main appeal of the proposed method. However, we think that one possibility would be to use a method where, at each node, a decision to use either a PO or a non PO model to split the node could be based on an objective criterion that would automatically check the PO assumption. Such a hybrid method could potentially combine the best of both worlds but we leave its investigation for future work.

REFERENCES

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.

Cloninger, C. R. (1987). Neurogenetic Adaptive Mechanisms in Alcoholism. *Science* **236**, 410-416.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society B* **34**, 187-202.

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.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. New York.

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.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random Survival Forests. *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.

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

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.

Mâsse, L. C. and Tremblay, R. E. (1997). Behavior of Boys in Kindergarten and the Onset of Substance Use During Adolescence. *Archives of General Psychiatry* **54**, 62-68.

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.

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

cial 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.

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.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press, New York.

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

Su, X., Wang, M. and Fan, J. (2004). Maximum Likelihood Regression Trees. *Journal of Computational and Graphical Statistics* **13**, 586-598.

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.

Zhang, H. (1998). Classification Trees for Multiple Binary Responses. *Journal of the American Statistical Association* **93**, 180-193.

APPENDIX

A1 – Maximum likelihood estimates of the hazards, the survival probabilities and the probabilities of event for the intercept only model (4):

For a general data set $(\tau_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, for which $\tau_i \in \{1, \ldots, K\}$, let $n_0(j)$ and $n_1(j)$ be the number of individuals for which $\tau_i = j$ and $\delta_i = 0$ and 1 respectively. For $j = 1, \ldots, K$, the number of individuals at risk at time j is

$$r(j) = \begin{cases} n & \text{if } j = 1 \\ n - \sum_{k=1}^{j-1}(n_0(k) + n_1(k)) & \text{if } j > 1. \end{cases}$$

The maximum likelihood estimates can be recursively defined as follows:

$$\hat{h}(j) = \begin{cases} n_1(j)/r(j) & \text{if } r(j) > 0 \\ 0 & \text{if } r(j) = 0, \end{cases}$$

$$\hat{S}(j) = \hat{S}(j-1)(1 - \hat{h}(j))$$

and

$$\hat{\pi}(j) = \begin{cases} 1 - \hat{S}(j) & \text{if } j = 1 \\ \hat{S}(j-1) - \hat{S}(j) & \text{if } j > 1. \end{cases}$$

A2 – Technical details for the pruning and selection of a single tree in Section (2.3):

The technical details to implement the pruning and bootstrap corection for the selection of the final tree are briefly described here. The reader is referred to LeBlanc

and Crowley (1993) or Fan *et al.* (2006) for a more complete description.

Firstly, here is how to get the sequence of pruned subtrees. We start with the large tree A_0. For any internal node t of A_0, let $A_0(t)$ denote the branch with t as its root node. The weakest link of A_0 is the node such that $G(A_0(t))/|W(A_0(t))|$ is the smallest among all internal nodes of A_0. Define α_1 to be this smallest value. Let A_1 be the subtree of A_0 after pruning off the branch with the weakest link as its root node. The same process of finding the weakest link and pruning the corresponding branch is applied to A_1 to produce a value α_2 and a new subtree A_2. Thus, starting from the large tree A_0, the pruning process can proceed recursively until the tree with only one node (root node), A_M, is reached. From LeBlanc and Crowley (1993), A_m is the smallest subtree that maximizes G_α, see (10), for all α in the interval $[\alpha_m, \alpha_{m+1})$. The geometric mean

$$\alpha'_m = \sqrt{\alpha_m \alpha_{m+1}} \tag{13}$$

is then assigned as the representative α value for A_m.

Secondly, here is how to get a bootstrap corrected value for $G(A)$ in (9). Draw B bootstrap samples from the original sample. Using the b^{th} sample, S_b, grow and prune a large tree. For $m = 1, \ldots, M$, let $A_b(\alpha'_m)$ denote the pruned subtree, in the sequence of subtrees constructed with S_b, corresponding to α'_m of (13), i.e., corresponding to the representative α value for the tree A_m from the original sequence of pruned subtrees for the original sample. The bootstrap corrected value of $G(A_m)$ is then

$$G^c(A_m) = G(A_m) - \frac{1}{B}\sum_{b=1}^{B}(G(A_b(\alpha'_m); S_b, S_b) - G(A_b(\alpha'_m); S_b, S))$$

where $G(A; S_1, S_2)$ denotes the value of $G(A)$ when the tree is built with the sample S_1 and the splitting statistics are recomputed with the sample S_2.

Paper 2: Discrete-Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data

Imad Bou-Hamad, Denis Larocque and Hatem Ben-Ameur

Department of Management Sciences
HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine,
Montréal, QC, Canada H3T 2A7

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.

Keywords: Bankruptcy data; Discrete-time survival analysis; Survival forests; Time-varying covariate.

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 for-

mulation 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

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

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

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

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

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

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

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

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) (Ghosh, 2006). 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, 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.

3.2. Results

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

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

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

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.

Parameter	Original Ratios	Transformed 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**

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.

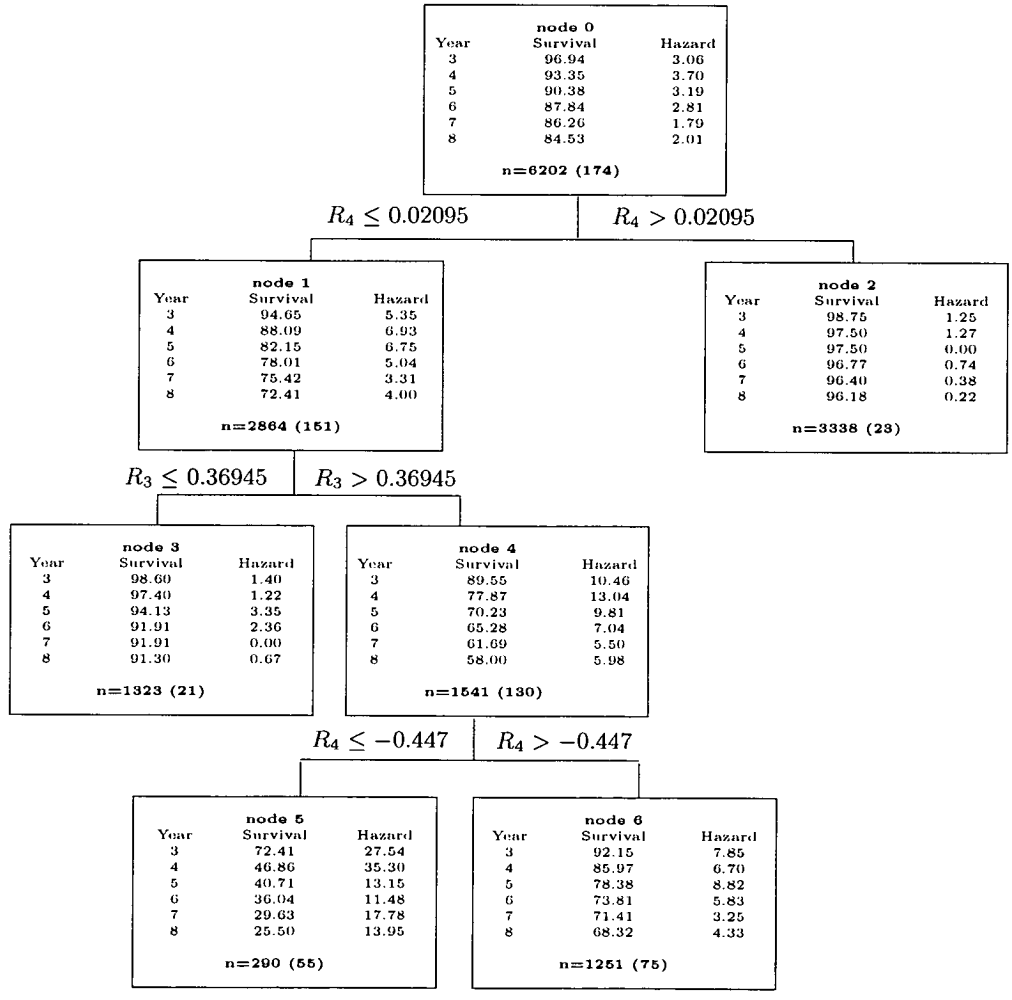

Figure 1: A single survival tree for the bankruptcy data

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.

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 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.

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.

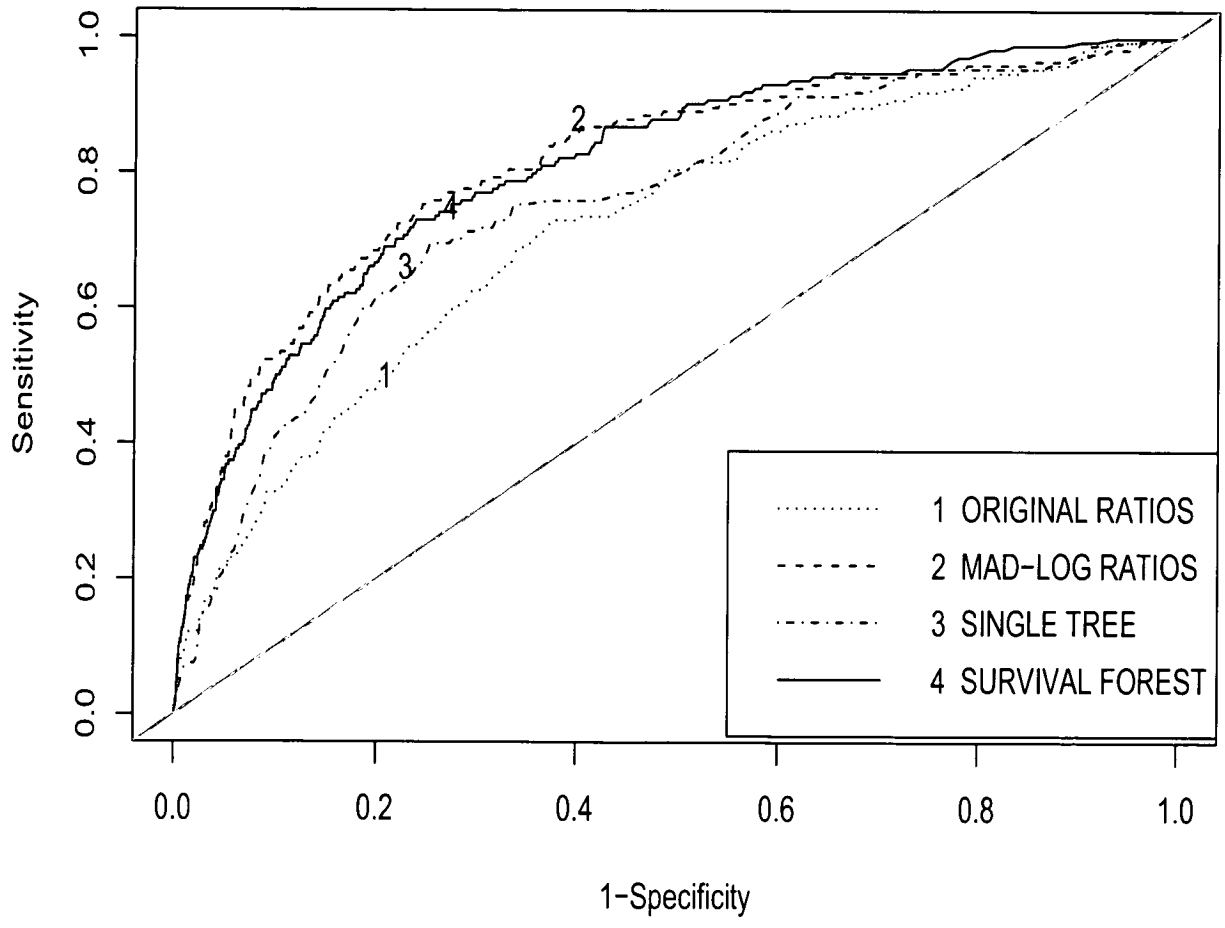

Figure 2: ROC curves for the out-of-sample risk estimates with the bankruptcy data
Four models are represented: DTPO models with the original and MAD-log ratios, a single tree and a survival forest with 100 trees.

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 inter-

pretability 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. *Canadian Journal of Statistics* **37**, 17-32.

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.

Ghosh, A. (2006). The IPOs Phenomenon in the 1990s. *The Social Science Journal* **43**, 487-495.

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.

Paper 3 : A Review of Survival Trees

Imad Bou-Hamad, Denis Larocque and Hatem Ben-Ameur

Department of Management Sciences
HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine,
Montréal, QC, Canada H3T 2A7

ABSTRACT

This paper presents a non–technical account of the developments of 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, to more recent work that are focusing on more complex models, more specialized methods and more specific problems including multivariate data, the use of time–varying covariates, survival data on a discrete scale, and ensemble methods applied to survival trees.

Keywords : Survival trees ; CART ; Time–varying covariate ; Right–censored data ; Discrete–time ; Ensemble methods ; Time–varying effect ; Bagging ; Survival forest.

1. INTRODUCTION

Studies involving time–to–event data are numerous and arise from all areas of research. The presence of censored data (most often right–censored) characterizes most of these studies and many statistical methods were developed to deal specifically with this problem. The Cox proportional hazard regression model and its extensions are very popular models to study survival variables with censoring. Survival trees are popular nonparametric alternatives to (semi) parametric models. They offer great flexibility for data exploration and can naturally group subjects according to their survival behavior based on their covariates. Prognostic groups can then be derived more easily from survival trees as opposed to regression type models. Moreover, survival trees are ideal candidates for combination through an ensemble method leading to very powerful predictive tools.

The development of survival trees followed a steady growth from the mid 80's up to the mid 90's, where the goal was mainly to extend the 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, the research moved into many different directions. One direction was to treat more complex situations like multivariate and correlated survival data. Another direction was to study the use of ensemble methods with survival trees. Also, another one was to deal with specific topics related to survival studies like time–varying covariates and time–to–event variables measured on a discrete scale.

Survival trees have been applied in numerous studies but, until now, mainly in the medical area. As a matter of fact, the vast majority of the articles that will be discussed in this review include examples of applications in various medical studies.

The rest of this section describes the basic tree methodology and the survival data setup. Section 2 focuses on the basic survival tree methodologies. In Section 3,

the more recent developments and extensions are presented. Finally, some concluding remarks are given in Section 4.

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 data without censoring. They were introduced by Morgan and Sonquist (1963) but really became popular in the 80's due in great part to the development of the CART (Classification and Regression Tree) paradigm described in the monograph by Breiman, Friedman, Olshen and Stone (1984). The reader is assumed to be familiar with the basic ideas and terminology around tree–based methods as only a brief description is provided here. The basic idea of a tree is to recursively partition the covariates space 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 the most often used measure for 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 to 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. 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 while the node proportions of each values of a categorical outcome 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 lead 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 covariates values are available at time 0 for each subject. Thus, only the baseline values of a time–varying covariate is typically used. The inclusion of the multiple values of time–varying covariates will be discussed in Section 3.3. Multivariate and correlated survival data will be the topic of Section 3.1.

2. SURVIVAL TREE BUILDING METHODS

The early idea of using a tree–structured data analysis for censored data can be traced back to Ciampi, Bush, Gospodarowicz and Till (1981) and Marubini, Morabito and Valsecchi (1983). However, the first paper that contained all the elements of what would become survival trees is the one by Gordon and Olshen (1985). In this section, we are presenting separately the splitting criteria and the final tree selection methods

that were proposed over the years. We are also presenting some variants and related methods and the few studies that compared some tree–building procedures.

2.1. Splitting Criteria

In this subsection, we will focus only on the different splitting criteria that were proposed.

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 at most one finite point. Although this particular splitting criterion did not gain much popularity, it laid ground to the work that followed. Indeed, Gordon and Olshen (1985) mention the possibility to use the logrank statistic or a parametric likelihood ratio statistic to measure the "distance" between the two children nodes and these ideas were used widely in the work that followed.

Ciampi, Thiffault, Nakache and Asselain (1986) proposed to use the logrank statistic to compare the two groups formed by the children nodes. The retained split was the one with the largest significant test statistic value. The use of the logrank test leads to a split which assures that the median survival times in the two children nodes are separated best. Ciampi, Chang, Hogg and Mckinney (1987) proposed a general formulation based around 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 is, the more dissimilar the two nodes are. They discuss more specifically about two possibilities : an exponential model and a Cox proportional hazards model. Hence, this approach relies on the assumptions related to the chosen model. For instance, with the Cox model, the proportional hazards assumption implies that the hazard function in the right node

is proportional to the one in the left node. Davis and Anderson (1989) also used a splitting criterion based on an exponential model log–likelihood which is equivalent to the LRS dissimilarity measure under the exponential model. Ciampi, Hogg, McKinney, and Thiffault (1988) and Ciampi, Thiffault and Sagman (1989) continued in the same direction and mention the possibility of using the logrank and Wilcoxon–Gehan statistics as dissimilarity measures and to use the Kolmogorov–Smirnov statistic to compare the survival curves of the two nodes. Segal (1988) also adopted 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 like the logrank and Wilcoxon–Gehan statistics. LeBlanc and Crowley (1993) are also using the logrank statistic as splitting criterion but they introduced a new method of pruning and selection of a final tree built around a measure of split–complexity (see Section 2.2.1).

In their discussion, Therneau, Grambsch and Fleming (1990) mentioned 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 softwares could be used directly with the modified outcome. Keles and Segal (2002) provided 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 which prevents the direct use of a 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) proposed 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 selecting as the splitting covariate the one 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 hazards model full likelihood and a Poisson model likelihood, Leblanc and Crowley (1992) proposed 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 algorithm in R or Splus.

Zhang (1995) proposed 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) proposed a unified strategy for building trees with censored data. Their approach is based on defining 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 uses the node variance as the impurity measure, Jin, Lu, Stone and Black (2004) proposed a splitting rule based on the variance of the survival time. But since the mean and variance survival times are affected by the censored observations, they proposed to compute the variance by using a restricted time limit.

Finally, Cho and Hong (2008) proposed to use 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 of a tree building algorithm is to decide when to stop splitting and hence select a specific tree as the final model. Too large trees will tend to overfit the data and not generalize well to the population of interest while too small trees might miss important characteristics of the relationship between the covariates and the outcome. There are basically two approaches for 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 branches off. The second one is a forward method which uses a built–in stopping rule to decide when to stop splitting a node further.

2.2.1. 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) used 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 optimally subtrees for any value of α. The split–complexity method is also used in Fan, Su, Levine, Nunn and LeBlanc (2006) and Bou-Hamad, Larocque, Ben-Ameur, Mâsse, Vitaro and Tremblay (2009).

2.2.2. 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

are : test set, cross–validation, bootstrap, AIC/BIC and graphical methods ("kink" in the curve or elbow method).

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) proposed two methods. The aim of both of them 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 to use 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 a 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, Chang, Hogg and Mckinney (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 to recompute $ll(T)$ 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. Similarly to a scree plot in a principal components analysis, such a plot usually have 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). A drawback of graphical methods is the subjectivity associated with them. Negassa, Ciampi, Abrahamowicz, Shapiro and Boivin (2000) proposed an automatic elbow detection method and applied it with an AIC criterion, as above but computed on the same sample as the one that built the tree.

2.2.3. 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 have been proposed to avoid possible selection bias in the choice of the covariate (Schittgen, 1999 and Lausen, Hothorn, Bretz and Schumacher, 2004).

At the same time, adjusted p–value can be used to regulate the tree building procedure, acting as a stopping criterion to decide 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 standardize two–sample logrank statistic as the splitting criterion, Lausen et al. (2004) proposed such a method which adjusts both 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).

2.3. Some variants and related methods

The RECPAM (Recursive Partition and Amalgamation) method introduced in Ciampi et al. (1988) allows an additional feature compared to a classical tree; see Ciampi, Negassa and Lou (1995) for a complete description. The method share the basic characteristics of regular trees in the sense that it builds a large tree, prunes it and 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 as a pruning and selection algorithm 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 covariates space may not necessarily be that of a tree since terminal nodes that are far away may end up grouped together. But it may bring down the number of groups to a more easily interpretable size. In their data example, Fan, Su, Levine, Nunn and LeBlanc (2006) used 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 group together terminal nodes which are similar with respect to the survival profiles was proposed in Tsai, Chen, Chen, Balch, Thompson and Soong (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 hazards models with piecewise constant relative risk functions. By adapting the ideas of Logical Analysis of Data or LAD (Hammer and Bonates, 2006), Kronek and Reddy (2008) proposed the method LASD (Logical Analysis of Survival Data) that automatically detects good patterns of covariates to predict the survival function. Finally, Su and Tsai (2005) proposed a hybrid approach to augment a Cox proportional hazards model through a tree structure.

2.4. Comparison of methods

A large scale simulation study to compare many 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 under the RECPAM framework, Negassa, Ciampi, Abrahamowicz, Shapiro and Boivin (2000, 2005) 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 concluded that none among theses approaches exhibited a uniformly superior performance over the different scenarios. They also proposed a two–stage method, where cross–validation is used in the first stage followed by the elbow approach, which performed well in the simulation.

A large scale comparison of many splitting criteria has also yet to appear. Some limited results appear in Radespiel-Tröger, Rabenstein, Schneider and Lausen (2003) and Radespiel-Tröger, Gefeller, Rabenstein and Hothorn (2006). In addition to a real data set, 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 was used in the first paper. Comparing many splitting methods, the authors concluded that the adjusted and unadjusted logrank statistic splitting with pruning, the exponential loss splitting with pruning and the adjusted logrank statistic

splitting without pruning have the best performance . Radespiel-Tröger et al. (2006) used bootstrap samples from a real data set to perform the simulation study. Their results showed that the adjusted logrank statistic splitting without pruning gave the best results.

3. EXTENSIONS OF THE BASIC METHODS

The last section presented the developments of 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 90's. This section will present these developments in a thematic fashion. Extensions to multivariate and correlated data will be presented first, followed by the use of ensemble methods with a survival tree as the base model. Finally, specialized topics like time–varying covariates and time–to–event variables measured on a discrete scale will be presented.

3.1. Multivariate and Correlated Data

A natural extension of the 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 to handle correlated survival outcomes and both of them have been adapted to build survival trees.

Su and Fan (2004) and Gao, Manatunga and Chen (2004) used 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 :

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

where h_0 is an unspecified baseline hazard function and w_i is a frailty term for cluster i that follows some known distribution. The gamma distribution was assumed in both papers. Su and Fan (2004) built a splitting criterion based on an integrated log–likelihood while Gao, Manatunga and Chen (2004) defined theirs through the standardized estimate of the splitting variable parameter obtained from a profile log–likelihood.

Fan, Su, Levine, Nunn and LeBlanc (2006) used the marginal approach where the dependence structure is left unspecified. Instead, their splitting criterion is based on 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.

3.2. 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 the predicting 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 predictors at each node. Bagging is then just the special case where all predictors are retained at each node.

Dannegger (2000) and Benner (2002) described applications of bagging with survival trees but the first two systematic studies appeared in 2004.

Ishawaran, Blackstone, Pothier and Lauer (2004) proposed to built a forest of relative risk trees using the tree building method introduced in LeBlanc and Crowley (1992) 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, Lausen, Benner and Radespiel-Tröger (2004) proposed a general bagging method for an arbitrary tree growing algorithm but used 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 together from each tree all the observations, from the bootstrap sample used to build the tree, that fell into the same terminal node as $\mathbf{x}$. Then they 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.

Hothorn, Bühlman, Dudoit, Molinaro, van der Laan (2006) proposed a random forest method to build a survival ensemble for the log–survival time. Their approach is based on the general Molinaro et al. (2004) framework (see Section 2.1). 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 investigated a gradient boosting algorithm where a tree can act as the base learner but they studied instead the use of componentwise 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 proposed boosting algorithm with different base learners.

Ishwaran, Kogalur, Blackstone and Lauer (2008) introduced a general random forest method coupled with a new algorithm for imputing missing values. They investigated 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 estimate of the cumulative hazard function at each node are averaged. Their method is implemented in the R package randomSurvivalForest.

Eckel, Pfahlberg, Gefeller and Hothorn (2008) compared proportionnal hazards models, survival forests and a bundling method with a data set of melanoma patients. The bundling method combines the Cox model with a tree by adding the linear predictor of a Cox model as an additional predictor, thus expanding the candidate splits. The final predictions are obtained from aggregated trees. Their conclusion was that the three methods were on par for this data set.

3.3. Specific Topics : Time–Varying Effects and Covariates, and Discrete–Time Survival Outcome

Almost all survival tree methods were developed under the basic setup described in section 1.2 and did not include time–varying effects nor time–varying covariates. Moreover, no method specifically adapted to discrete–time survival data were proposed

until very recently.

Given that time–varying covariates are common in practice, only the difficulties associated with their use can explain the sparsity of the literature on tree based methods about 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 with time. For instance, if the values of a time–varying covariate of an individual are regressed against 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 that dealt with time–varying covariates with survival trees were the ones by Bacchetti and Segal (1995) and Huang, Chen and Soong (1998).

The solution proposed by Bacchetti and Segal (1995) was to allow the decomposition of each subject into pseudo–subjects defined through the splitting rules of the tree. 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 splitted into two pseudo–subjects that could be splitted apart 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) used modified two–sample test statistics that can accommodate left–truncated data.

Huang, Chen and Soong (1998) used a similar approach in which subjects can

be splitted 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) methods used only time independent covariates but these were allowed to have time–varying effects. With their method, the tree is used only to find splits for the time variable in order to locate the time values where the effect change 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, Larocque, Ben-Ameur, Mâsse, Vitaro and Tremblay (2009) proposed 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 (2009) generalized this approach to be able to incorporate time–varying covariates. This was achieved by allowing subjects to be splitted across different nodes depending on the time period as in Bacchetti and Segal (1995). Hence, this method allows simultaneously time–varying effects and time–varying covariates. These two papers also investigated the use of bagging and random forests which produce aggregate estimations of the discrete conditional risk and survival functions.

4. CONCLUSION

This review shows that survival trees have been and are still a very active area of research. Many methods were 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 were studied and the development of ensemble methods renewed the interest about tree based methods in general and survival trees in particular.

However, there are many topics that still need further research. For instance, it is still unclear which survival tree method should be recommended as there are only few articles that tried to systematically compare the many different approaches. Much work is also needed about the use of time–varying covariates and about how to incorporate time–varying effects. Finally, the interpretation of covariates effects with ensemble of trees in general is still mainly unsolved and should attract future research. In the context of survival trees, one further difficulty arises when time–varying effects are included.

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*

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.

Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2009). Discrete-Time Survival Trees and Forests with Time-Varying Covariates : Application to Bankruptcy Data. Submitted.

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.

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.

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.

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., 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.

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.

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

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.

Ishwaran H., Kogalur U. B., Blackstone E. H. and Lauer M. S. (2008). Random Survival Forests.*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.

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

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., Schumacher, M. (2004). Assessment of Optimal Selected Prognostic Factors. *Biometrical Journal* **46**, 364-374.

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.

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 n 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.

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.

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.

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.

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

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.

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.

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

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.

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.

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

CONCLUSION GÉNÉRALE

Nous avons divisé la thèse en trois articles. Dans le premier article, une nouvelle méthodologie pour construire un arbre à temps discret a été présentée. Le critère de séparation utilisé dans la méthode proposée est équivalent au critère de l'entropie, utilisé pour une variable réponse catégorielle, en l'absence de censure. Par conséquent, la nouvelle méthode pourrait être considérée comme une extension de la méthode d'arbre de classification pour des données censurées à droite. De plus, les covariables peuvent avoir des effets qui varient dans le temps. Comme le critère de séparation peut être évalué rapidement grâce à une formule explicite, le temps de calcul ne pose pas problème et des techniques d'agrégation d'arbres, telles le bagging et les forêts aléatoires, peuvent facilement être utilisées. Des simulations ont montré que la nouvelle méthode performe bien. D'autre part, elle a été illustrée avec un exemple avec des données sur le tabagisme chez les adolescents. Cette méthodologie est surtout recommandée dans le cas où le nombre de périodes observées est limité.

La principale motivation du deuxième article était l'étude des facteurs reliés à la faillite. De nombreux travaux se sont intéressés à la modélisation de la faillite. Les premières méthodes utilisaient seulement une seule période dans le temps mais depuis quelques années, des approches utilisant plusieurs périodes, par l'entremise de modèle d'analyse de survie, ont vu le jour. Les méthodes d'analyse de survie à temps discret sont le plus souvent utilisées étant donné que le statut de l'entreprise ainsi que les covariables utilisées (souvent des ratios financiers) sont évalués annuellement. Ainsi, les covariables varient dans le temps. C'est pourquoi le deuxième article propose une extension de la méthode de base du premier article afin de pouvoir inclure de telles covariables. Les résultats de l'exemple avec les données de faillite montrent qu'une forêt de survie construite avec la nouvelle méthode d'arbre performe mieux qu'un modèle paramétrique de base qui utilise les ratios financiers tel quels et performe de manière équivalente à un modèle paramétrique qui utilise une transformation particulière des

ratios. L'avantage d'une forêt de survie est qu'elle ne requiert pas de choix de la part de l'analyste contrairement aux modèles paramétriques qui eux nécessitent des choix non -triviaux concernant la manière d'inclure les covariables (choix des transformations, choix des interactions etc.) Cependant, l'interprétation est plus difficile avec une forêt de survie. Si l'interprétation est primordiale, une forêt de survie demeure tout de même une référence pour comparer la performance d'autres modèles plus interprétables.

Finalement, nous avons présenté dans le troisième article une étude exhaustive des méthodes d'arbres de survie. Cette étude contribue à la littérature en mettant à jour la revue présentée dans Leblanc Crowley (1995). Nous nous sommes concentrés sur les éléments fondamentaux comme les critères de séparation et les méthodes de sélection d'un arbre final. En plus, nous avons décrit les nouveaux développements qui sont apparus depuis Leblanc et Crowley (1995), tels les méthodes pour données de survie multivariés, l'utilisation de méthodes d'ensemble avec les arbres de survie ainsi que certains sujets précis comme les covariables et les effets variant dans le temps.

BIBLIOGRAPHIE

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

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.

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

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.

Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2009). Discrete-Time Survival Trees and Forests with Time-Varying Covariates: Application to Bankruptcy Data. Submitted.

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.

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 Compari-

son 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.

Cloninger, C. R. (1987). Neurogenetic Adaptive Mechanisms in Alcoholism. *Science* **236**, 410-416.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society B* **34**, 187-202.

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.

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.

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

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., 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.

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

Ghosh, A. (2006). The IPOs Phenomenon in the 1990s. *The Social Science Journal* **43**, 487-495.

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

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

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. New York.

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.

Ishwaran H., Kogalur U. B., Blackstone E. H. and Lauer M. S. (2008). Random Survival Forests.*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.

Keles, S. and Segal, M. R. (2002). Residual-Based Tree–Structured Survival Analysis

Statistics in Medicine **21**, 313-326.

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., Schumacher, M. (2004). Assessment of Optimal Selected Prognostic Factors. *Biometrical Journal* **46**, 364-374.

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.

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 n Medicine* **2**, 295-303.

Mâsse, L. C. and Tremblay, R. E. (1997). Behavior of Boys in Kindergarten and the Onset of Substance Use During Adolescence. *Archives of General Psychiatry* **54**, 62-68.

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.

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.

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.

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

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

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

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.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press, New York.

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 Ap-

proach Based on Frailty Models. *Biometrics* **60**, 93-99.

Su, X., Wang, M. and Fan, J. (2004). Maximum Likelihood Regression Trees. *Journal of Computational and Graphical Statistics* **13**, 586-598.

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.

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.

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.

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

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