Tree-augmented Cox proportional hazards models

XIAOGANG SU*
Department of Statistics and Actuarial Science,
University of Central Florida, Orlando, FL 32816, USA
xsu@pegasus.cc.ucf.edu

CHIH-LING TSAI
Graduate School of Management, University of California, Davis, Davis, CA 95616, USA, and
Guanghua School of Management, Peking University, Beijing, P.R. China

SUMMARY
We study a hybrid model that combines Cox proportional hazards regression with tree-structured modeling. The main idea is to use step functions, provided by a tree structure, to ‘augment’ Cox (1972) proportional hazards models. The proposed model not only provides a natural assessment of the adequacy of the Cox proportional hazards model but also improves its model fitting without loss of interpretability. Both simulations and an empirical example are provided to illustrate the use of the proposed method.

Keywords: BIC; Censored survival times; Cox proportional hazards models; Survival trees.

1. INTRODUCTION
The Cox (1972) proportional hazards (hereafter referred to as PH) model is of central importance in the analysis of censored failure time data. It has a semiparametric form, but efficient inference can be made based on a partial likelihood approach (Cox, 1975). In addition, the PH assumption provides a reasonable empirical approximation to the association between covariates and censored failure times in applications. In more recent years, tree-based methods (Morgan and Sonquist, 1963; Breiman et al., 1984) have been extended to survival analysis by many authors. A brief and not exhaustive survey includes Gordon and Olshen (1985), Ciapmi et al. (1986), Segal (1988), Davis and Anderson (1989), Therneau et al. (1990), Loh (1991), LeBlanc and Crowley (1992, 1993), Ahn and Loh (1994), and Xu and Adak (2002). As a nonparametric tool in nature, a tree-based method approximates the hazard function with piecewise constants by recursively partitioning the predictor space into disjoint regions.

In this paper, we study a hybrid model that combines the Cox PH regression with tree modeling. One primary motivation of this research stems from the observation that, in many aspects, the Cox PH regression and tree-structured modeling complement each other. First, the Cox PH model postulates that the logarithm of the hazard rate is linearly related to the covariates and it is capable of modeling their linear effects. Comparatively, it takes a large tree structure to model a simple linear effect. Nevertheless, when nonlinear effects or complex interactions are involved, the Cox PH model may encounter problems whereas tree methods are often found quite useful in these situations. Furthermore, while tree-based methods show great efficiency in coping with categorical variables, the Cox PH model defines dummy...
variables to include categorical predictors and may result in a massive model form when many categorical predictors are involved or a categorical predictor has many levels. On the other hand, both modeling methods provide meaningful interpretation. Therefore, combining them can yield a hybrid model that provides a more efficient way to model survival data and improves fit without loss of interpretability.

Our motivation is further stimulated by the question of how to combine them effectively so that the hybrid model can utilize the strengths of both methods. Ideally, the hybrid model would address the issue of Cox PH model selection in the context of recursive partitioning. However, one intrinsic difficulty is that the computational burden inevitably becomes overwhelming. To circumvent this problem, we propose to augment a fitted Cox PH model with a survival tree structure. The main idea is to first select a ‘best’ Cox PH model and then use an augmentation tree structure to approximate what is leftover. The resulting model provides a natural adequacy checking for the functional form specification in the ‘best’ Cox PH model.

The remainder of the paper is organized in the following manner. Section 2 presents in detail the augmentation tree procedure which constitutes the hybrid model. Section 3 provides Monte Carlo studies to investigate the proposed tree procedure. Section 4 illustrates its use by analyzing an empirical example, and Section 5 concludes the paper with a discussion.

2. TREE-AUGMENTED COX PH MODELS

Let $F_i$ and $C_i$ be the failure time and the censoring time of the $i$th case, respectively. The observed data consist of $\{(X_i, \delta_i, z_i): i = 1, 2, \ldots, n\}$, where $X_i = \min(F_i, C_i)$ is the $i$th observed failure time; $\delta_i = 1_{\{F_i \leq C_i\}}$ is the indicator of whether the $i$th case is a failure or censored; and $z_i = (z_{i1}, z_{i2}, \ldots, z_{ip}) \in \mathcal{R}^p$ is a $p$-dimensional covariate vector for the $i$th case. We assume noninformative censoring in the sense that $F_i$ and $C_i$ are independent given the covariate vector $z_i$.

The Cox (1972) proportional hazards model is defined as

$$\lambda(t | z) = \lambda_0(t) \exp\{g(z)\},$$

where $\lambda_0(t)$ is the unspecified baseline hazard function of time and $g(\cdot)$ is a function of $z$. In practice, $g(\cdot)$ is frequently specified, even though this risks its misspecification, as a linear combination of selected components, say, $z_0$, of $z$ and hence the working model is

$$\lambda(t | z) = \lambda_0(t) \exp\{\beta'z^0\},$$

where $z^0 \in \mathcal{R}^q$ is a $q$-dimensional vector selected from $z$, which may have incorporated terms such as dummy variables, interactions, and transformations, and $\beta = (\beta_1, \ldots, \beta_q)' \in \mathcal{R}^q$ is a vector of unknown regression parameters.

Suppose that model (2.2) is the ‘best’ fitted model that one has come up with using conventional Cox model fitting and selection techniques. To check its adequacy and improve its fit, we use a regression tree structure to augment model (2.2). This constitutes the following hybrid model

$$\lambda(t | z) = \lambda_0(t) \exp\{\beta'z^0 + \gamma'z^{(T)}\},$$

where $z^{(T)}$ is a dummy vector defined by a tree structure $T$ such that $z^{(T)}_{ij} = 1$, if the $i$th observation belongs to the $j$th terminal node of $T$, and 0, otherwise, and $\gamma$ is the corresponding parameter vector. Hereafter, we refer to $T$ as an augmentation tree for model (2.2), which, heuristically speaking, is designed to provide an approximation using piecewise constant step functions to the leftover part $g(z) - \beta'z^0$. For a given augmentation tree structure $T$, let $\bar{T}$ denote the set of all terminal nodes in $T$ and $|\cdot|$ denote cardinality. Then, $\gamma = (\gamma_1, \ldots, \gamma_{|\bar{T}|})' \in \mathcal{R}^{|\bar{T}|}$ subject to $\sum_{j=1}^{|\bar{T}|} \gamma_j = 0$ for identifiability concern. The new model (2.3) has a semiparametric specification, although its form is as simple as parametric.
Combining the strengths from both linear regression and tree modeling results in four main advantages for the hybrid model. First, it captures global linear patterns with the parametric part $\beta'z^0$ and, at the same time, picks up local properties such as nonlinearity, thresholds, and interactions with the augmentation tree $T$. By doing so, more predictive power can be achieved. Second, the hybrid model has the flexibility to handle complex data structures more efficiently. For example, continuous covariates can be handled by either linear modeling or the tree method while categorical variables with many levels can be left to the tree structure. Third, meaningful interpretations can be extracted by combining the analysis results from both the linear fitting and the tree structure. In general, the augmentation tree $T$ tends to be of smaller sizes than an ordinary survival tree and hence more interpretable. Fourth, the augmentation tree structure not only provides insightful diagnostic information about the Cox model but also gives valuable clues about how to amend the model specification. Toward this end, a test of $H_0: \gamma = 0$ can be used to assess the adequacy of model (2.2). On the other hand, if one fails to develop a nontrivial augmentation tree, then it is likely that the original Cox model provides a good fit.

In the next three subsections, we present a procedure for constructing the augmentation tree $T$ by making iterative adjustment to model (2.2). The proposed procedure adopts the pruning idea of classification and regression trees (CART; Breiman et al., 1984), which consists of first growing a large initial tree, then truncating it to obtain a nested sequence of subtrees, and finally selecting the best tree size.

### 2.1 Growing a large tree

To split the data, we consider the following model

$$\lambda(t \mid z) = \lambda_0(t) \exp(\beta'z^0 + \nu_1z_1 \leq c).$$

(2.4)

The indicator function $1_{\{z_1 \leq c\}}$, which equals 1 if the $j$th continuous predictor $z_j$ is less than or equal to a cutpoint $c$ and 0 otherwise, includes a binary split $s$ of the data. When $z_j$ is nominal with levels in $C = \{c_1, \ldots, c_k\}$, the form of $1_{\{z_j \in A\}}$ for any subset $A \subset C$ is considered. The resulting partial log-likelihood of model (2.4) given a permissible split $s : z_j \leq c$ is

$$l(\beta, \nu \mid s) = \sum_{i=1}^{n} \delta_i \left[ \beta'z_i^0 + \nu_1z_{ij} \leq c \right] - \log \left( \sum_{k=1}^{n} 1_{\{X_k \geq x_i\}} \exp(\beta'z_k^0 + \nu_1z_{kj} \leq c) \right).$$

(2.5)

Finding the best split, say $s^*$, corresponds to the problem of estimating the change-point covariate effect. This can be done by maximizing, over all permissible splits, a test statistic for assessing $H_0: \nu = 0$ in model (2.4). Here, we consider the score test (Rao, 1973), which is asymptotically equivalent to the likelihood ratio test but does not require evaluation of the model under the alternative hypothesis. The score test in this setting, often called the \textit{covariate-adjusted logrank} test, is given by

$$S(\nu \mid s) = U^2(\hat{\beta}_0, 0)/\hat{\Sigma},$$

(2.6)

where

$$U(\hat{\beta}_0, 0) = \left[ \frac{\partial}{\partial \nu} l(\beta, \nu \mid s) \right]_{(\beta = \hat{\beta}_0, \nu = 0)} = \left\{ \sum_{i=1}^{n} \delta_i r_i \right\}_{(\beta = \hat{\beta}_0, \nu = 0)},$$

(2.7)

with

$$r_i = 1_{\{z_{ij} \leq c\}} - \frac{\sum_{k=1}^{n} 1_{\{X_k \geq x_i\}} 1_{\{z_{ij} \leq c\}} \exp(\beta'z_k^0)}{\sum_{k=1}^{n} 1_{\{X_k \geq x_i\}} \exp(\beta'z_k^0)}.$$
being the Schonfeld’s (1982) residuals from model (2.4) associated with the dummy variable \(1_{\{z_{ij} \leq c\}}\) and \(\hat{\beta}_0\) being the maximum partial likelihood (MPL) estimator of \(\beta\) under \(H_0\), and \(\hat{\Sigma}\) is a consistent estimator of the variance of \(U(\beta_0, 0)\). The detailed expressions of \(\hat{\Sigma}\) can be shown to be

\[
\hat{\Sigma} = \{I_{vv} - I_{v\beta}I_{\beta\beta}^{-1}I_{v\beta}\}_{(\beta = \hat{\beta}_0, v = 0)},
\]

where

\[
I_{vv} = \sum_{i=1}^{n} \delta_i \sum_{k=1}^{n} w_{ki} 1\{z_{ij} \leq c\} \sum_{k=1}^{n} w_{k1} 1\{z_{ij} > c\},
\]

\[
I_{v\beta} = I_{\beta\beta} = \sum_{i=1}^{n} \delta_i \sum_{k=1}^{n} w_{ki} (1\{z_{ij} \leq c\} \sum_{l=1}^{n} w_{li} - \sum_{l=1}^{n} w_{li} 1\{z_{ij} \leq c\}) z_{k0}^0,
\]

\[
I_{\beta\beta} = \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{k=1}^{n} w_{ki} z_{k0}^{0\otimes 2}}{\sum_{k=1}^{n} w_{ki}} - \left( \frac{\sum_{k=1}^{n} w_{ki} z_{k0}^0}{\sum_{k=1}^{n} w_{ki}} \right)^{\otimes 2} \right\},
\]

\[w_{ki} = 1_{\{X_i > X_j\}} \exp(\hat{\beta}_0^T z_i^0),\text{ and } a^{\otimes 2} = aa'\text{ for a vector } a.\]

Further discussion on computing the score test statistic \(S(v|s)\) can be found in Therneau and Grambsch (Section 3.4, 2000).

For a given split \(s\), the null distribution of \(S(v|s)\) is asymptotically \(\chi^2\) with one degree of freedom. The best split \(s^*\) corresponds to the one that yields the maximum value of score test statistics among all permissible splits, i.e. \(S(v|s^*) = \max_s S(v|s)\). The data are then partitioned into two child nodes according to the best split \(s^*\), and subsequently the same splitting procedure is applied to partition either child node. Recursively doing so yields a large initial tree \(T^0\).

**Remark 1** LeBlanc and Crowley (1995) developed a weighted least squares method to approximate the score test statistic in their proposed additive step function models, which facilitates fast updating of the splitting statistics for neighboring cutpoints. Their method can be directly adopted to serve in our setting. In addition, the robust score tests derived from the robust inference studied by Tsiatis et al. (1985), Lin and Wei (1989), Kong and Slud (1997), etc. can also be used as splitting statistics.

### 2.2 Pruning

The final tree model could be any subtree of \(T^0\). To narrow down its choices, we follow CART’s (Breiman et al., 1984) pruning idea to iteratively truncate the ‘weakest’ link of the large initial tree \(T^0\). Since we employ the maximized score statistics criterion to grow the tree, which provides a goodness-of-split measure for each link, it is natural to adopt the split-complexity pruning algorithm of LeBlanc and Crowley (1993). We now briefly describe this procedure.

Start with the large initial tree \(T^0\). For any link or internal node \(h\) of \(T^0\), let \(T_h\) denote the branch whose root node is \(h\) and \(I_h\) be the set of internal nodes in \(T_h\). Define

\[
g(h) = \frac{\sum_{i \in I_h} S(i)}{|I_h|},
\]

where \(S(i)\) represents the maximum score test statistic associated with internal node \(i\). Then the ‘weakest link’ \(h^*\) in \(T_0\) is the internal node such that \(g(h^*) = \min_{h \in I_0} g(h)\), where \(I_0\) denotes the set of all internal nodes in \(T^0\). Now let \(T^1\) be the subtree after pruning off \(h^*\), namely, \(T^1 = T^0 - T_{h^*}\). Then apply the same procedure to truncate \(T^1\) to obtain the subtree \(T^2\). Continuing this way leads to a decreasing sequence of
subtrees $T^M \prec \cdots \prec T^1 \prec T^0$, where $T^M$ is the root node and the notation $\prec$ means ‘is a subtree of’. Note that the quantities of $S(i)$ have been computed when growing $T_0$, which renders the above pruning procedure computationally very fast.

### 2.3 Tree size selection

To select the best-sized tree from the nested sequence, it is convenient to employ the widely used Bayesian information criterion (BIC, Schwarz, 1978). Since tree-based methods are not very often recommended for small samples, we suppose that the sample size is large or at least moderate. To proceed, we randomly divide the data $\mathcal{L}$ into two groups: the learning (or training) sample $\mathcal{L}_1$ and the test (or validation) sample $\mathcal{L}_2$ with a ratio of the sample sizes $n_1/n_2$ approximately 2:1. A large tree is then grown and pruned using the learning sample $\mathcal{L}_1$ so that a decreasing sequence of subtrees $\{T^m: 0 \leq m \leq M\}$ is obtained.

Next, send the test sample down each subtree $T^m$, and then compute its validated partial log-likelihood (up to a constant)

$$l^{(T^m)}(\hat{\beta}, \hat{\gamma}) \propto \sum_{i: (X_i, \delta_i=1, z_i) \in \mathcal{L}_2} \left[ \hat{\gamma}' z_i^{(T^m)} - \log \left\{ \sum_{k: (X_k, \delta_k, z_k) \in \mathcal{L}_2} 1_{\{X_k \geq X_i\}} \exp \left( \hat{\beta}' z_k^0 + \hat{\gamma}' z_k^{(T^m)} \right) \right\} \right],$$

where $(\hat{\beta}, \hat{\gamma})$ are the MPL estimates of $(\beta, \gamma)$ computed from the learning sample $\mathcal{L}_1$. Since the tree structure $T^m$, constructed using the learning sample $\mathcal{L}_1$, can be treated as known when applied to the independent test sample $\mathcal{L}_2$, the total number of parameters associated with the partial likelihood of model (2.3) is hence $(q + |T^m| - 1)$. The BIC for model (2.3) is given by, up to a constant,

$$\text{BIC}^{(T^m)} \propto -2l^{(T^m)}(\hat{\beta}, \hat{\gamma}) + \log(n)(q + |\bar{T}^m| - 1) \quad \propto -2l^{(T^m)}(\hat{\beta}, \hat{\gamma}) + \log(n)|T^m|.$$  

For censored failure time data, Volinsky and Raftery (2000) suggested replacing $\log(n)$ by $\log(\sum \delta_i)$, i.e. the logarithm of the total number of observed failures. We adopt their suggestion to compute BIC. A smaller BIC$^{(T)}$ corresponds to a more favorable tree structure $T$. Thus, the best-sized subtree $T^*$ is the one that provides the minimum BIC, i.e. $\text{BIC}^{(T^*)} = \min_{0 \leq m \leq M} \text{BIC}^{(T^m)}$.

Once the best tree structure $T^*$ is identified, we pool the learning sample and the test sample together and then refit model (2.3) with all the data.

**Remark 2** To construct survival trees, Therneau et al. (1990) suggested a residual-based approach, as further explored by Keles and Segal (2002), by applying a regression tree procedure such as CART (Breiman et al., 1984) directly on the null martingale residuals. Analogously, a natural alternative approach to construct the augmentation tree is to apply a regression tree procedure directly on the residuals computed from model (2.2).

### 3. Simulations

This section contains simulations designed to assess the performance of the proposed augmentation tree procedure and compares it with the residual-based approach. Various types of residuals have been developed for Cox proportion hazards models. Following Therneau et al.’s (1990) original suggestion, martingale residuals are natural choices. However, since Therneau and Grambsch (2000) show that martingale residuals perform poorly with respect to model selection and that deviance residuals afford superior results, we shall pursue comparisons using deviance residuals as well.
Tree-augmented Cox proportional hazards models

Two sets of Monte Carlo studies are presented. The first investigates the splitting schemes, and the second evaluates different tree methods in detecting data structure.

3.1 Comparing splitting methods

We first compare three different splitting methods: the one proposed in Section 2.1 (Method I), and two residual-based approaches addressed in Remark 2 (Method II and Method III). In residual-based approaches, the best split is selected by maximizing the reduction in the corrected sum of squares of residuals due to a split. Method II uses martingale residuals while Method III employs deviance residuals.

The data were generated from the following two models:

Model A: \( \lambda(t) = \exp(z_1 + 2 \times 1_{z_2 \leq 0.5}) \) and
Model B: \( \lambda(t) = \exp(z_1 + 2 \times 1_{z_2 \leq 0.5}) \),

where covariates \( z_1 \) and \( z_2 \) were simulated from a discrete uniform distribution over values \( \{1/50, \ldots, 50/50\} \). The difference between Models A and B is that the linear part and the tree part are confounded in Model B, but not in Model A. Two censoring rates, 0 and 50%, were also considered.

For each set of 100 generated observations, three splitting methods were applied to identify the best cutpoint, and for comparative convenience all started with the ‘best’ Cox model \( \lambda(t) = \exp(\beta z_1) \). Figure 1 presents the relative frequencies of selected cutpoints based on 200 runs. The bar at the cutpoint \(-0.5\) corresponds to the case when the splitting variable is wrongly chosen (e.g. the covariate \( z_1 \) is chosen to split the data from Model A, which results in a spurious split). It can be seen that Method I performs comparatively better than the two residual-based methods in the nonconfounded Model A, and outperforms them by a great deal in the confounded Model B. While the performances of the three methods all deteriorate with heavier censoring, Method I also tends to be less sensitive. In terms of comparison between two residual-based methods, using deviance residuals (Method III) seems to offer slightly better results than using martingale residuals (Method II). However, neither of them gives satisfactory results for confounded cases.

3.2 Tree structure detection

Next, we compare three tree procedures in detecting the data structure: the one proposed in Sections 2.1–2.3 (Method I*), the martingale residual-based tree procedure addressed in Remark 2 (Method II*), and the deviance residual-based tree procedure (Method III*).

We generated data from the following five models:

Model A*: \( \lambda(t) = \exp(2z_1 + 2z_2); \)
Model B*: \( \lambda(t) = \exp(2z_1 + 2z_2 + 4 \times 1_{z_1 \leq 0.5 \land z_2 \leq 0.5}); \)
Model C*: \( \lambda(t) = \exp(2z_1 + 2z_2 + 4 \times 1_{z_3 \leq 0.5 \land z_4 \leq 0.5}); \)
Model D*: \( \lambda(t) = \exp(4z_1 + 4z_2 + 2 \times 1_{z_1 \leq 0.5 \land z_2 \leq 0.5}); \) and
Model E*: \( \lambda(t) = \exp(4z_1 + 4z_2 + 2 \times 1_{z_3 \leq 0.5 \land z_4 \leq 0.5}); \)

where \( z_1, \ldots, z_4 \) were again simulated from a discrete uniform distribution over \( \{1/50, \ldots, 50/50\} \). Model A* is the null model containing only the linear part. Models C* and E* differ from models B* and D* according to whether the same set of covariates are involved in both the linear part and the tree structure. Models B* and C* have a relatively stronger tree signal than Models D* and E*. Each data set contains 600 observed failure times: 400 for the learning sample and 200 for the test sample. Two censoring rates were considered: no censoring and 50% censoring. For each combination, 100 realizations were conducted.
Fig. 1. The comparison of three different splitting methods: Method I (the score test), Method II [maximum reduction of corrected sum of squares (CSS) in martingale residuals of Model (2.2)], and Method III (maximum reduction of CSS in deviance residuals).
Table 1. The comparison of three tree methods: the proposed tree procedure (I*) and CART procedures based on martingale residuals (II*) and deviance residuals (III*). The frequencies of the final tree sizes correctly selected by each method are marked in boldface.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tree method</th>
<th>Final tree size</th>
<th>Variable selection</th>
<th>Final tree size</th>
<th>Variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1   2   3   4   5</td>
<td></td>
<td>1   2   3   4   5</td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>I*</td>
<td>99  1  0  0  0  0</td>
<td>99</td>
<td>97  2  1  0  0  0</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>II*</td>
<td>94  6  0  0  0  0</td>
<td>94</td>
<td>94  6  0  0  0  0</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>III*</td>
<td>96  3  1  0  0  0</td>
<td>96</td>
<td>95  5  0  0  0  0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>(0% censoring)</td>
<td></td>
<td></td>
<td>(50% censoring)</td>
<td></td>
</tr>
<tr>
<td>B'</td>
<td>I*</td>
<td>1   0  98  1  0  0</td>
<td>99</td>
<td>1   0  91  7  1  0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>II*</td>
<td>3   0  3   18 11  65</td>
<td>79</td>
<td>15  0  15  23 19  28</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>III*</td>
<td>0   0  2   37 11  50</td>
<td>86</td>
<td>14  0  46  26 8  6</td>
<td>75</td>
</tr>
<tr>
<td>C'</td>
<td>I*</td>
<td>0   0  100  0  0  0</td>
<td>100</td>
<td>0   0  90  6  4  0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>II*</td>
<td>0   0  90  9  0  1</td>
<td>95</td>
<td>2   1  88  6  2  1</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>III*</td>
<td>0   0  87  9  3  1</td>
<td>94</td>
<td>1   2  85  9  2  1</td>
<td>92</td>
</tr>
<tr>
<td>D'</td>
<td>I*</td>
<td>3   0  89  8  0  0</td>
<td>95</td>
<td>32  1  58  6  1  2</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>II*</td>
<td>15  2  5   16 21  41</td>
<td>46</td>
<td>61  3  3   11 8  14</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>III*</td>
<td>6   1  17  27 26  23</td>
<td>75</td>
<td>72  4  6   8  6  4</td>
<td>17</td>
</tr>
<tr>
<td>E'</td>
<td>I*</td>
<td>0   0  93  4  3  0</td>
<td>100</td>
<td>0   0  85  7  8  0</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>II*</td>
<td>0   0  90  5  3  2</td>
<td>97</td>
<td>6   4  81  4  5  0</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>III*</td>
<td>0   0  87  7  6  0</td>
<td>98</td>
<td>5   6  80  5  2  2</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 1 presents the frequencies of the final tree sizes (i.e. number of terminal nodes) selected by three proposed tree methods. Note that an ideal augmentation tree structure should have only one terminal node for data generated from Model A' and three for Models B' through E'. Furthermore, to address the variable selection issue (e.g. a good augmentation tree for Model B' or D' should involve both \( z_1 \) and \( z_2 \), but neither \( z_3 \) nor \( z_4 \)), we counted frequencies when a final tree was split according to exactly the desired variables.

Table 1 shows that the proposed procedure, Method I*, does fairly well in detecting true augmentation tree structures. Note that the results for the plain Cox linear model, i.e. Model A', roughly correspond to the size issue in statistical hypothesis testing. If we treat tree augmentation as if it is a testing procedure for lack-of-fit, then Table 1 shows that the ‘empirical sizes’ (i.e. the relative frequency when a nontrivial augmentation tree is wrongly developed for Model A') can be computed as \( 1 - 99/100 = 1\% \) when no censoring is involved and \( 1 - 97/100 = 3\% \) for the 50% censoring case. They stay well within the nominal level in both cases. In terms of comparison with the residual-based procedures, all three methods provide comparable results when the linear part does not confound with the tree structure. When they are confounded, Method I* substantially outperforms the two residual-based approaches. Furthermore, all methods deteriorate with heavier censoring and weaker tree signals as expected. However, Methods II* and III* fail more rapidly than Method I*. It is also interesting to note that Method III* using deviance residuals offers improved results than Method II* using martingale residuals in most cases considered here.

An implementation of the proposed augmentation tree method was done using SPLUS (version 6) on a 1.8 GHz personal computer with 1.00 GB RAM. The codes are available from the authors. For the simulations presented in Table 1, each run took an average of 14.68 seconds for Method I* and 2.12 seconds for the residual-based approach. So computationally Method I* is feasible, although the residual-based methods are more efficient. In conclusion, the residual-based methods are naturally adapted.
from conventional approaches and give quicker results in situations where computational efficiency is of great concern. However, we would like to recommend Method I* in routine use for the superiority that it offers in efficacy.

4. WESTERN COLLABORATIVE GROUP STUDY DATA

The Western Collaborative Group Study (WCGS) is a prospective cardiovascular epidemiology study. In 1960 through 1961, 3154 healthy middle-aged white males, who were free of coronary heart disease (CHD) and cancer, were drawn from 10 large Californian corporations and followed up for various health outcomes such as incidence of CHD, cancer, and death. See Ragland and Brand (1988) for a more detailed description of the study design. Many research papers have been published on this study, most of which are focused on the prevalence or risk of developing CHD. There are only a few, e.g. Carmelli et al. (1997), concerning cancer mortalities in WCGS. In this example, we pursue this latter concern.

After a 33-year follow-up till 1993, 405 deaths due to cancer were known. There are also 927 participants who died prior to 1993. But information on the causes of these deaths were unavailable, which could be or could not be because of cancer, we excluded these records from our analysis. Table 2 lists part of the baseline characteristics that were collected from the WCGS. After further excluding 180 records that contain missing values, we have 2047 observations remaining. We shall focus on this data set.

Two covariates, hostility and behpatt, are particularly worth noting. The variable behpatt is an indicator of Type A pattern behavior. The Type A pattern behavior is characterized by traits such as impatience, aggressiveness, a sense of time urgency, and the desire to achieve recognition and advancement. People exhibiting Type A behavior seem to find themselves in various high-pressured scenarios. Researchers have broken down the Type A behavior construct in order to study its many subcomponents. One of its major subcomponents is hostility, denoted by hostility. The ‘hostility’ part is characterized by a tendency to react to unpleasant situations with responses that reflect anger, frustration, irritation, and disgust. See e.g. Schneiderman et al. (1989) for more discussions. The measurements of hostility and behpatt in WCGS were done by means of a structured interview using two kinds of information, answers given to interview questions and the individual’s behavior during the interview. There has been compelling evidence (see, e.g. Rosenman et al., 1975) linking Type A behavior pattern and hostility to various aspects of CHD. But few cancer studies take behavioral measures into consideration. Therefore, one question of primary interest is whether a person’s behavior is associated with his risk of cancer death.

Carmelli et al. (1997) and Zhang and Singer (1999) analyzed the data using both Cox proportional hazards modeling and survival trees. The Cox proportional hazards model that they developed includes eversmok, age, chol, and wcr as predictors. We refer to this model as the ‘best’ fitted Cox model (2.2) and name it as Model M1, whose fit is given in Table 3. Note that our data set is slightly different from

| Table 2. Nine baseline covariates and their descriptions in the WCGS data |
|-----------------------------|-----------------------------|
| eversmok                    | Smoking habits (0 = never smoked; 1 = smoked) |
| age                         | Age in years                |
| educ                        | Education (1 = hs; 2 = college; 3 = graduate school) |
| sbp                         | Systolic blood pressure    |
| chol                        | Cholesterol                |
| hostlty                     | Hostility score            |
| behpatt                     | Behavior pattern (1 = Type A; 0 = others) |
| bmi                         | Body mass index            |
| wcr                         | Waist-to-calf ratio        |
Table 3. The model fittings (based on the entire WCGS data set) and their corresponding log-likelihood scores of the ‘best’ Cox Model M₁ and two tree-augmented Cox models, M₂ and M₃

| Estimate | SE  | z Value | 2 × Pr(>|z|) |
|----------|-----|---------|-------------|
| Model M₁: log-likelihood −2,607.67 |
| eversmok | 0.8047 | 0.1539 | 5.23 | <0.0001 |
| age      | 0.1206 | 0.0095 | 12.75 | <0.0001 |
| chol     | 0.0034 | 0.0013 | 2.69 | 0.0071 |
| wcr      | 0.7697 | 0.2688 | 2.86 | 0.0042 |
| Model M₂: log-likelihood −2,597.97 |
| eversmok | 0.7627 | 0.1545 | 4.94 | <0.0001 |
| age      | 0.1193 | 0.0095 | 12.54 | <0.0001 |
| chol     | 0.0033 | 0.0013 | 2.62 | 0.0087 |
| wcr      | -0.0959 | 0.3955 | -0.24 | 0.8085 |
| node 112 | 0.3679 | 0.1220 | 3.02 | 0.0026 |
| node 12  | 0.7601 | 0.1892 | 4.02 | <0.0001 |
| Model M₃: log-likelihood −2,598.00 |
| eversmok | 0.7617 | 0.1544 | 4.93 | <0.0001 |
| age      | 0.1188 | 0.0094 | 12.71 | <0.0001 |
| chol     | 0.0033 | 0.0013 | 2.61 | 0.0091 |
| node 112 | 0.3673 | 0.1220 | 3.01 | 0.0026 |
| node 12  | 0.7294 | 0.1407 | 5.19 | <0.0001 |

Their due to some minor modifications and updating. Besides, the 927 deaths with unknown causes were treated as censored (i.e. irrelevant to cancer) in their analysis.

Next, we apply the augmentation tree method to boost Model M₁. To proceed, we divide the whole data into three subsets: 1000 randomly selected observations for the learning sample, 500 observations for the test sample, and the rest 547 observations forming another independent validation sample. The validation sample set will be used to confirm the tree-augmented Cox model.

The learning sample was used to construct a large initial tree. After pruning, we sent the test sample down, and applied BIC to determine the final tree size. Figure 2 plots the final tree structure, which has three terminal nodes. Its internal nodes (i.e. nodes 1 and 11) are denoted by circles, and its terminal nodes (i.e. nodes 111, 112, and 12) by rectangles. The first split is according to whether the waist-to-calf ratio...
Fig. 3. The Kaplan–Meier survival curves for the WCGS data computed from three terminal nodes in the augmentation tree.

(wcr), a measure of obesity, exceeds 2.6 or not. Subsequently, for subjects with wcr $\leq 2.6$, who are not overweighted, their survival rates further differ according to their two behavior patterns. It can be found that 12.83% (153 out of 815) of healthy middle-aged men exhibiting Type A behavior died of cancer with a mean survival time of 9962 days, while 18.77% (121 out of 943) of those exhibiting other types of behaviors died of cancer with a mean survival time of 10349 days. The difference in mean survival times is about 1 year. In addition, it is a common practice in survival trees to plot the Kaplan–Meier survival estimates for all terminal nodes, which is presented in Figure 4. It shows striking differences among the three terminal nodes. The unadjusted logrank statistics between every pair of them, not reported here, all yield $p$-values of essentially 0.

Since a nontrivial augmentation tree structure has been constructed, the adequacy of the original Cox model becomes questionable. This lack-of-fit can be further confirmed using the test of $H_0: \gamma = 0$. The likelihood ratio test based on the validation sample yields a value of 10.436 and corresponds to a $p$-value of 0.0054 referring to the $\chi^2_{(2)}$ distribution, which clearly indicates that the pattern recognized by the augmentation tree is nonnegligible. Note that if the test were based on the learning sample or the test sample, then the result would be overoptimistic because of the optimally selected cutpoints. However, using an independent validation data set circumvents such a difficulty.

Finally, pooling the whole data together, we brought the identified tree structure into the Cox PH model by incorporating two dummy variables (labeled as nodes 112 and 12). The resulting tree-augmented
Cox model (Model M_2) is summarized in Table 3. Note that, interestingly, the effect of wcr becomes quite insignificant after incorporating the threshold on wcr. Hence, we deleted wcr and refit the tree-augmented Cox model, which yields Model M_3 in Table 3 and clearly provides a more parsimonious fit. In terms of interpretation, Model M_3 shows that, in addition to age, smoking status, and cholesterol level, obesity is associated with higher risk of cancer death among healthy middle-aged males. For those without obesity, their risk of cancer mortality is also affected by their behavioral patterns. In particular, the hazard ratio between Type A and non-Type A behavior groups, after adjusting for other covariates, is exp(0.3673) = 1.444. Note that neither of the two behavioral measures was significantly present in the original Cox Model M_1.

In conclusion, the tree-augmented Cox model not only assessed and remedied the deficiencies of the classical Cox model M_1 but also offered new insights into cancer death that were not uncovered from previous analysis. Nevertheless, the extent of their validity warrants further scientific investigation.

**Remark 3** It is worth noting that the adequacy test illustrated above resembles grouping-based goodness-of-fit tests, which has been widely studied for the Cox PH models in the literature. In particular, Gronesby and Borgan (1996) proposed an overall goodness-of-fit $\chi^2$ test by grouping martingale residuals, which is, as shown by May and Hosmer (1998), algebraically identical to one obtained from adding group indicator variables to the model and testing the hypothesis that the coefficients associated with the dummy variables are zero via the score test. One main shortcoming of the grouping-based tests is that their effectiveness and conclusion are sensitive to the number of groups and how the groups are defined. However, in our procedure, grouping is optimized through recursive partitioning. The tree-based augmentation not only makes the grouping procedure automatic and objective but also enhances the effectiveness of the related test, i.e. $H_0: \gamma = 0$. Because of the very adaptive nature of tree, how to obtain an ‘honest’ $p$-value is an issue here. We have employed an independent validation sample to avoid the overoptimism. When the sample size is small and the validation sample is unavailable, ad hoc approaches include the Bonferroni-type adjustment and permutation-based tests, which warrant further research efforts.

5. Discussion

We proposed a hybrid model combining Cox PH models with trees. We chose to amend an ordinary Cox model with an augmentation survival tree, which provides a feasible way of exploiting the merits of both methods. Our procedure starts with and centers around a carefully built Cox PH model. The resulting augmentation tree structure provides a natural assessment of the adequacy for the Cox PH model and automatically remedies its deficiencies without loss of interpretability. To our knowledge, this is the first time that survival trees have been used for this purpose. Because of the wide popularity of Cox PH models, we believe that the proposed method can make use of survival trees more extensively by attaching them naturally to Cox PH models. Another important reason for using trees as an augmentation tool can be attributed to their instability. One main criticism on recursive partitioning is its inherent instability on the ground that the tree structure may vary dramatically from sample to sample, which deters trees from being the sole modeling approach in many applications. Instead, using them as a supplemental method reasonably reduces the influence of an instable tree structure on both statistical modeling and interpretation. Moreover, since the hierarchical tree structure automatically classifies observations into disjoint groups, the proposed method can also find its applications in situations where prognosis or diagnosis rules are to be based on a group of readily available predictors whereas the classification needs to be adjusted for another group of uneasily accessed covariates.

To further explore the usefulness of this method, we identify two avenues for future research: first, consider different slopes of $z^0$ across terminal nodes of the augmentation tree and allow for splits on the
time scale (Xu and Adak, 2002); second, apply the tree augmentation approach to obtain other diagnostic measures for assessing the appropriateness of Cox models.

ACKNOWLEDGMENTS

The authors would like to thank the editor and one anonymous referee for their very constructive comments. In particular, the approach using deviance residuals was suggested by the referee. The Western Collaborative Group Study was funded by National Institute on Aging grant No. AG09341 (Gary E. Swan, Principal Investigator). We also would like to thank Dr. Lisa M. Jack for offering the data set and Ruth Krasnow for her efforts in assembling the data file.

REFERENCES


[Received April 27, 2004; first revision January 12, 2005; second revision February 24, 2005; accepted for publication March 1, 2005]