Unbiased Recursive Partitioning: A Conditional Inference Framework

Torsten Hothorn, Kurt Hornik, Achim Zeileis

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 8
July 2004

http://statistik.wu-wien.ac.at/

Unbiased Recursive Partitioning: A Conditional Inference Framework

Torsten Hothorn
Friedrich-Alexander-Universität
Erlangen-Nürnberg

Kurt Hornik
Wirtschaftsuniversität Wien

Achim Zeileis
Wirtschaftsuniversität Wien

Abstract

Recursive binary partitioning is a popular tool for regression analysis. Two fundamental problems of exhaustive search procedures usually applied to fit such models have been known for a long time: Overfitting and a selection bias towards covariates with many possible splits or missing values. While pruning procedures are able to solve the overfitting problem, the variable selection bias still seriously effects the interpretability of tree-structured regression models. For some special cases unbiased procedures have been suggested, however lacking a common theoretical foundation. We propose a unified framework for recursive partitioning which embeds tree-structured regression models into a well defined theory of conditional inference procedures. Stopping criteria based on multiple test procedures are implemented and it is shown that the predictive performance of the resulting trees is as good as the performance of established exhaustive search procedures. It turns out that the partitions and therefore the models induced by both approaches are structurally different, indicating the need for an unbiased variable selection. The methodology presented here is applicable to all kinds of regression problems, including nominal, ordinal, numeric, censored as well as multivariate response variables and arbitrary measurement scales of the covariates. Data from studies on animal abundance, glaucoma classification, node positive breast cancer and mammography experience are re-analyzed.

Keywords: permutation tests, variable selection, multiple testing, ordinal regression trees, multivariate regression trees.

1. Introduction

With their seminal work on automated interaction detection (AID), Morgan and Sonquist (1963) introduced a class of simple regression models for prediction and explanation nowadays known as ‘recursive partitioning’ or ‘trees’. Many variants and extensions have been published in the last 40 years, the majority of which are special cases of a simple two-stage algorithm: First partition the observations by univariate splits in a recursive way and second fit a constant model in each cell of the resulting partition. The most popular implementations of such algorithms are ‘CART’ (Breiman, Friedman, Olshen, and Stone 1984) and ‘C4.5’ (Quinlan 1993). Not unlike AID, both perform an exhaustive search over all possible splits maximizing an information measure of node impurity selecting the covariate showing the best split. This approach has two fundamental problems: Overfitting and a selection bias towards covariates with many possible splits. Within the exhaustive search framework, pruning procedures, mostly based on some form of cross-validation, are necessary to restrict the number of cells in the resulting partitions in order to avoid overfitting problems. While pruning is successful in selecting the right-sized tree, the interpretation of the trees is affected by the biased variable selection. This bias is induced by maximizing a splitting criterion over all possible splits simultaneously and was identified as a problem by many researchers (e.g., Kass 1980; Segal 1988; Breiman *et al.* 1984, p. 42). The nature of the variable selection problem under different circumstances has been studied intensively (White and Liu 1994; Jensen and Cohen 2000; Shih 2004) and Kim and Loh (2001) argue that exhaustive search methods are biased towards variables with many missing values as well. With this article we enter at the point where White and Liu (1994) demand for "[...] a *statistical* approach [to recursive partitioning] which takes into account the *distributional* properties of the measures." We present a unified framework embedding recursive binary partitioning into the well defined theory of permutation tests developed by Strasser and Weber (1999). The conditional distribution of statistics measuring the association between responses and covariates is the basis for an unbiased selection among covariates measured at different scales. Moreover, multiple test procedures are applied to determine when no significant association between any of the covariates and the response can be stated and the recursion needs to stop. We show that such statistically motivated stopping criteria implemented via hypothesis tests lead to regression models whose predictive performance is equivalent to the performance of optimally pruned trees obtained by well established exhaustive search methods, therefore offering an intuitive and computationally efficient solution to the overfitting problem.

The development of the framework presented here was inspired by various attempts to solve both the overfitting and variable selection problem. The χ^2 automated interaction detection algorithm ('CHAID' Kass 1980) is the first approach based on statistical significance tests for contingency tables. Unbiasedness is achieved by a separation of variable selection and splitting procedure. The significance of the association between a nominal response and one of the covariates is investigated by a χ^2 test and the covariate with highest association is selected for splitting. For a binary response variable and continuous covariates, Rounds (1980) proposes an algorithm utilizing the distribution of the Kolmogorov-Smirnov statistic. A series of papers for nominal and continuous responses starts with 'FACT' (Loh and Vanichsetakul 1988), where covariates are selected within an analysis of variance framework treating a nominal response as independent variable. Loh and Shih (1997) extend those results in order to construct unbiased binary trees for nominal responses ('QUEST'). The case of a continuous response is handled by Loh (2002), called 'GUIDE'. Although those variants do not suffer a variable selection bias, they are not generally applicable and require to categorize continuous covariates prior to modeling (CHAID) or treat covariates as dependent variables when the association to the response is measured by ANOVA or Kolmogorov-Smirnov statistics.

Permutation (or randomization) tests offer a rather simple non-parametric solution to those problems: The validity of a split can be assessed by permuting the responses under the null hypothesis of independence between covariates and response variable (Jensen and Cohen 2000; LeBlanc and Crowley 1993; Frank and Witten 1998). In practical applications however, the computation of the permutation distribution in a recursive algorithm renders this approach rather burdensome. For certain special cases, computationally feasible solutions have been suggested: The asymptotic distribution of maximally selected rank statistics (Lausen and Schumacher 1992) can be applied to correct the bias of exhaustive search recursive partitioning for continuous and censored responses (see Lausen, Hothorn, Bretz, and Schumacher 2004, and the references therein). An approximation to the distribution of the Gini criterion is given by Dobra and Gehrke (2001). However, lacking solutions for more general situations, these auspicious approaches are hardly ever applied and the majority of tree-structured regression models reported and interpreted in applied research papers is biased. The main reason is that computationally efficient solutions are available for special cases only. The framework presented in Section 3 is efficiently applicable to regression problems where both response and covariates can be measured at arbitrary scales, including nominal, ordinal, discrete and continuous as well as censored and multivariate variables. The treatment of special situations is explained in Section 4 and applications including animal abundance, glaucoma classification, node positive breast cancer and a questionnaire on mammography experience illustrate the methodology in Section 5. Finally, we show that recursive partitioning based on statistical criteria as introduced in this paper lead to regression models whose predictive performance is as good as the performance of optimally pruned trees by means of benchmark experiments.

2. Recursive Binary Partitioning

We focus on regression models describing the conditional distribution of a response variable $\mathbf{Y}$ given the status of m covariates by means of tree-structured recursive partitioning. The response $\mathbf{Y}$ from some sample space $\mathcal{Y}$ may be multivariate as well. The m covariates $\mathbf{X} = (X_1, \dots, X_m)$ are element of a sample space $\mathcal{X} = \mathcal{X}_1 \times \dots \mathcal{X}_m$. Both response variable and covariates may be measured at arbitrary scales. We assume that the conditional distribution $D(\mathbf{Y}|\mathbf{X})$ of the response $\mathbf{Y}$ given the covariates $\mathbf{X}$ depends on a function f of the covariates

$$D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y}|X_1, \dots, X_m) = D(\mathbf{Y}|f(X_1, \dots, X_m)),$$

where we restrict ourselves to partition based regression relationships, i.e. r disjoint cells $B_1, \dots, B_r$ partitioning the covariate space $\mathcal{X} = \bigcup_{k=1}^{r} B_k$. A model of the regression relationship is to be fitted based on a learning sample $\mathcal{L}_n$, i.e., a random sample of n independent and identically distributed observations, possibly with some covariates X_{ji} missing,

$$\mathcal{L}_n \quad = \quad \{(\mathbf{Y}_i, X_{1i}, \dots, X_{mi}); i = 1, \dots, n\}.$$

A generic algorithm for recursive binary partitioning for a given learning sample $\mathcal{L}_n$ can be formulated using non-negative integer valued case weights $\mathbf{w} = (w_1, \dots, w_n)$. Each node of a tree is represented by a vector of case weights having non-zero elements when the corresponding observations are element of the node and are zero otherwise. The following algorithm implements recursive binary partitioning:

1. For case weights $\mathbf{w}$ test the global null hypothesis of independence between any of the m covariates and the response. Stop if this hypothesis can not be rejected. Otherwise select the covariate X_{j^*} with strongest association to $\mathbf{Y}$.

2. Choose a set $A^* \subset \mathcal{X}_{j^*}$ in order to split $\mathcal{X}_{j^*}$ into two disjoint sets A^* and $\mathcal{X}_{j^*} \setminus A^*$. The case weights $\mathbf{w}_{\text{left}}$ and $\mathbf{w}_{\text{right}}$ determine the two subgroups with $w_{\text{left},i} = w_i I(X_{j^*i} \in A^*)$ and $w_{\text{right},i} = w_i I(X_{j^*i} \notin A^*)$ for all $i = 1, \dots, n$ ($I(\cdot)$ denotes the indicator function).

3. Recursively repeat steps 1 and 2 with modified case weights $\mathbf{w}_{\text{left}}$ and $\mathbf{w}_{\text{right}}$, respectively.

As we sketched in the introduction, the separation of variable selection and splitting procedure into steps 1 and 2 of the algorithm is the key for the construction of interpretable tree structures not suffering a systematic tendency towards covariates with many possible splits or many missing values. In addition, a statistically motivated and intuitive stopping criterion can be implemented: We stop when the global null hypothesis of independence between the response and any of the m covariates can not be rejected at a pre-specified nominal level α. The algorithm induces a partition $\{B_1, \dots, B_r\}$ of the covariate space $\mathcal{X}$, where each cell $B \in \{B_1, \dots, B_r\}$ is associated with a vector of case weights.

3. Recursive Partitioning by Conditional Inference

In the main part of this section we focus on step 1 of the generic algorithm. Unified tests for independence are constructed by means of the conditional distribution of linear statistics in the permutation test framework developed by Strasser and Weber (1999). The determination of the best binary split in one selected covariate and the handling of missing values is performed based on standardized linear statistics within the same framework as well.

Variable Selection and Stopping Criteria.

At step 1 of the generic algorithm given in Section 2 we face an independence problem. We need to decide whether there is any information about the response variable covered by any of the m covariates. In each node identified by case weights $\mathbf{w}$, the global hypothesis of independence is formulated in terms of the m partial hypotheses $H_0^j : D(\mathbf{Y}|X_j) = D(\mathbf{Y})$ with global null hypothesis $H_0 = \bigcap_{j=1}^m H_0^j$. When we are not able to reject H_0 at a pre-specified level α, we stop the recursion. If the global hypothesis can be rejected, we measure the association between $\mathbf{Y}$ and each of the covariates $X_j, j = 1, \dots, m$, by test statistics or P-values indicating the deviation from the partial hypotheses H_0^j.

For notational convenience and without loss of generality we assume that the case weights w_i are either zero or one. The symmetric group of all permutations of the elements of $(1, \dots, n)$ with corresponding case weights $w_i = 1$ is denoted by $S(\mathcal{L}_n, \mathbf{w})$. A more general notation is given in the Appendix. We measure the association between $\mathbf{Y}$ and $X_j, j = 1, \dots, m$, by linear statistics of the form

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = \mathrm{vec}\left(\sum_{i=1}^n w_i g_j(X_{ji}) h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^{p_j q} \tag{1}$$

where $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is a non-random transformation of the covariate X_j. The *influence function* $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ depends on the responses $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ in a permutation symmetric way. Section 4 explains how to chose g_j and h in different practical settings. A $p_j \times q$ matrix is converted into a $p_j q$ column vector by column-wise combination using the 'vec' operator.

The distribution of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ under H_0^j depends on the joint distribution of $\mathbf{Y}$ and X_j, which is unknown under almost all practical circumstances. At least under the null hypothesis one can dispose of this dependency by fixing the covariates and conditioning on all possible permutations of the responses. This principle leads to test procedures known as *permutation tests*. The conditional expectation $\mu_j \in \mathbb{R}^{p_j q}$ and covariance $\Sigma_j \in \mathbb{R}^{p_j q \times p_j q}$ of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ under H_0 given all permutations $\sigma \in S(\mathcal{L}_n, \mathbf{w})$ of the responses are derived by Strasser and Weber (1999):

$$\begin{aligned}
\mu_j &= \mathbb{E}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})|S(\mathcal{L}_n, \mathbf{w})) = \mathrm{vec}\left(\left(\sum_{i=1}^n w_i g_j(X_{ji})\right) \mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))^\top\right), \\
\Sigma_j &= \mathbb{V}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})|S(\mathcal{L}_n, \mathbf{w})) \\
&= \frac{\mathbf{w}.}{\mathbf{w}. - 1} \mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})^\top\right) \\
&- \frac{1}{\mathbf{w}. - 1} \mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji})\right) \otimes \left(\sum_i w_i g_j(X_{ji})\right)^\top
\end{aligned} \tag{2}$$

where $\mathbf{w}. = \sum_{i=1}^n w_i$ denotes the sum of the case weights, and $\otimes$ is the Kronecker product. The conditional expectation of the influence function is

$$\mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}.^{-1} \sum_i w_i h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) \in \mathbb{R}^q$$

with corresponding $q \times q$ covariance matrix

$$\begin{aligned}
\mathbb{V}(h|S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}.^{-1} \sum_i w_i &\left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) - \mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))\right) \\
&\left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) - \mathbb{E}(h|S(\mathcal{L}_n, \mathbf{w}))\right)^\top.
\end{aligned}$$

Having the conditional expectation and covariance at hand we are able to standardize a linear statistic $\mathbf{T} \in \mathbb{R}^{pq}$ of the form (1). Univariate test statistics c mapping an observed linear statistic $\mathbf{t} \in \mathbb{R}^{pq}$ into the real line can be of arbitrary form. An obvious choice is the maximum of the absolute values of the standardized linear statistic $c_{\max}(\mathbf{t}, \mu, \Sigma) = \max(|\mathbf{t} - \mu| / \mathrm{diag}(\Sigma)^{-1/2})$ utilizing the conditional expectation μ and covariance matrix Σ. The application of a quadratic form $c_{\mathrm{quad}}(\mathbf{t}, \mu, \Sigma) = (\mathbf{t} - \mu)\Sigma^+(\mathbf{t} - \mu)^\top$ is one alternative, although computational more expensive because the Moore-Penrose inverse Σ^+ of Σ is involved. It is important to note that the test statistics $c(\mathbf{t}_j, \mu_j, \Sigma_j), j = 1, \dots, m$, can not be directly compared in an unbiased way unless all of the covariates are measured at the same scale, i.e., $p_1 = p_j, j = 2, \dots, m$. In order to allow for an unbiased variable selection we need to switch to the P-value scale because P-values for the conditional distribution of test statistics $c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j)$ can be directly compared among covariates measured at different scales. In step 2 of the generic algorithm we select the covariate with minimum P-value, i.e., the covariate X_{j^*} with $j^* = \operatorname{argmin}_{j=1,\dots,m} P_j$, where

$$P_j = \mathbb{P}_{H_0^j}(c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j) \geq c(\mathbf{t}_j, \mu_j, \Sigma_j) | S(\mathcal{L}_n, \mathbf{w}))$$

denotes the P-value of the conditional test for H_0^j. So far, we have only addressed testing each partial hypothesis H_0^j, which is sufficient for an unbiased variable selection. A global test for H_0 required in step 1 can be constructed via an aggregation of the transformations $g_j, j = 1, \dots, m$, i.e., using a linear statistic of the form

$$\mathbf{T}(\mathcal{L}_n, \mathbf{w}) = \operatorname{vec}\left(\sum_{i=1}^{n} w_i \left(g_1(X_{1i})^\top, \dots, g_m(X_{mi})^\top\right)^\top h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n))^\top\right).$$

However, this approach is less attractive for learning samples with missing values. Universally applicable approaches are multiple test procedures based on $P_1, \dots, P_m$. Simple Bonferroni-adjusted P-values mP_j or a min-P-value resampling approach are just examples and we refer to the multiple testing literature (e.g., Westfall and Young 1993) for more advanced methods. We reject H_0 when the minimum of the adjusted P-values is less than a pre-specified nominal level α and otherwise stop the algorithm. In this sense, α may be seen as a unique parameter determining the size of the resulting trees.

The conditional distribution and thus the P-value of the statistics $c(\mathbf{t}, \mu, \Sigma)$ can be computed in several different ways. For some special forms of the linear statistic, the exact distribution of the test statistic is trackable. Conditional Monte-Carlo procedures can be used to approximate the exact distribution. Strasser and Weber (1999) proved (Theorem 2.3) that the conditional distribution of linear statistics $\mathbf{T}$ with conditional expectation μ and covariance Σ tends to a multivariate normal distribution with parameters μ and Σ as $n, s \to \infty$. Thus, the asymptotic conditional distribution of test statistics of the form $c_{\max}$ is normal and can be computed directly in the univariate case ($pq = 1$) or approximated by means of quasi-randomized Monte-Carlo procedures in the multivariate setting (Genz 1992). For quadratic forms c_{quad} which follow a χ^2 distribution with degrees of freedom given by the rank of Σ (Theorem 6.20 Rasch 1995), exact probabilities can be computed efficiently.

Splitting Criteria.

The same framework is utilized to find the optimal binary split in one selected covariate X_{j^*} in step 2 of the generic algorithm. The goodness of a split is evaluated by two-sample linear statistics which are special cases of the linear statistic (1). For all possible subsets A of the sample space $\mathcal{X}_{j^*}$ the linear statistic

$$\mathbf{T}_{j^*}^A(\mathcal{L}_n, \mathbf{w}) = \operatorname{vec}\left(\sum_{i=1}^{n} w_i I(X_{j^*i} \in A) h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^q$$

induces a two-sample statistic measuring the discrepancy between the samples $\{\mathbf{Y}_i | w_i > 0 \land X_{ji} \in A; i = 1, \dots, n\}$ and $\{\mathbf{Y}_i | w_i > 0 \land X_{ji} \not\in A; i = 1, \dots, n\}$. The conditional expectation $\mu_{j^*}^A$ and covariance $\Sigma_{j^*}^A$ can be computed by (2). The split A^* with a test statistic maximized over all possible subsets A is established:

$$A^* = \underset{A}{\operatorname{argmax}}\, c(\mathbf{t}_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A). \tag{3}$$

Note that we do not need to compute the distribution of $c(\mathbf{t}_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A)$ in step 2. In order to anticipate pathological splits one can restrict the number of possible subsets that are evaluated, for example by introducing restrictions on the sample size or the sum of the case weights in each of the two groups of observations induced by a possible split.

Missing Values and Surrogate Splits.

If an observation X_{ji} in covariate X_j is missing, we set the corresponding case weight w_i to zero for the computation of $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ and, if we would like to split in X_j, in $\mathbf{T}_j^A(\mathcal{L}_n, \mathbf{w})$ as well. Once a split A^* in X_j has been implemented, surrogate splits can be established by searching for a split leading to roughly the same division of the observations as the original split. One simply replaces the original response variable by a binary variable $I(X_{ji} \in A^*)$ coding the split and proceeds as described in the previous part.

4. Examples

Univariate Continuous or Discrete Regression.

For an univariate numeric response $\mathbf{Y} \in \mathbb{R}$, the most natural influence function is the identity $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = \mathbf{Y}_i$. In some cases a ranking of the observations may be appropriate: $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = \sum_{k=1}^n w_k I(\mathbf{Y}_k \leq \mathbf{Y}_i)$ for $i = 1, \dots, n$. Numeric covariates can be handled by the identity transformation $g_{ji}(x) = x$ (ranks are possible, too). Nominal covariates at levels $1, \dots, K$ are represented by $g_{ji}(k) = e_K(k)$, the unit vector of length K with kth element being equal to one. Due to this flexibility, special test procedures like the Spearman test, the Wilcoxon-Mann-Whitney test or the Kruskal-Wallis test and permutation tests based on ANOVA statistics or correlation coefficients are covered by this framework. Splits obtained from (3) maximize the absolute value of the standardized difference between two means of the values of the influence functions.

Censored Regression.

The influence function h may be chosen as Logrank or Savage scores taking censoring into account and one can proceed as for univariate continuous regression. This is essentially the approach first published by Segal (1988). An alternative is the weighting scheme suggested by Molinaro, Dudoit, and van der Laan (2004).

J-Class Classification.

The nominal response variable at levels $1, \dots, J$ is handled by influence functions $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = e_J(\mathbf{Y}_i)$. Note that for a nominal covariate X_j at levels $1, \dots, K$ with $g_{ji}(k) = e_K(k)$ the corresponding linear statistic $\mathbf{T}_j$ is a vectorized contingency table.

Ordinal Regression.

Ordinal response variables measured at J levels, and ordinal covariates measured at K levels, are associated with score vectors $\xi \in \mathbb{R}^J$ and $\gamma \in \mathbb{R}^K$, respectively. Those scores reflect the 'distances' between the levels: If the variable is derived from an underlying continuous variable, the scores can be chosen as the midpoints of the intervals defining the levels. The linear statistic is now a linear combination of the linear statistic $\mathbf{T}_j$ of the form

$$\mathbf{M}\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) \quad = \quad \mathrm{vec}\left(\sum_{i=1}^n w_i \gamma^\top g_j(X_{ji}) \left(\xi^\top h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)\right)^\top\right)$$

with $g_j(x) = e_K(x)$ and $h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) = e_J(\mathbf{Y}_i)$. If both response and covariate are ordinal, the matrix of coefficients is given by the Kronecker product of both score vectors $\mathbf{M} = \gamma \otimes \xi \in$

$\mathbb{R}^{1,KJ}$. In case the response is ordinal only, the matrix of coefficients $\mathbf{M}$ is a block matrix

$$\mathbf{M} = \left(\begin{array}{ccc|c|ccc} \xi_1 & & 0 & & \xi_q & & 0 \\ & \ddots & & \dots & & \ddots & \\ 0 & & \xi_1 & & 0 & & \xi_q \end{array} \right) \text{ or } \mathbf{M} = \operatorname{diag}(\gamma)$$

when one covariate is ordered but the response is not. For both $\mathbf{Y}$ and X_j being ordinal, the corresponding test is known as linear-by-linear association test (Agresti 2002).

Multivariate Regression.

For multivariate responses, the influence function is a combination of influence functions appropriate for any of the univariate response variables discussed in the previous paragraphs: i.e., indicators for multiple binary responses (Zhang 1998; Noh, Song, and Park 2004), Logrank or Savage scores for multiple failure times (for example tooth loss times Su and Fan 2004) and the original observations or a rank transformation for multivariate regression (De'ath 2002; Larsen and Speckman 2004).

5. Illustrations and Applications

In this section, we present regression problems which illustrate the potential fields of application of the methodology. Conditional inference trees based on c_{quad}-type test statistics using the identity influence function and asymptotic χ^2 distribution are applied. For the stopping criterion a simple Bonferroni correction with $\alpha = 0.05$ is used. Two additional restrictions are made: Nodes with less than 20 observations are not split and each split needs to send at least 1% of the observations into each of the two daughter nodes. The computations are based on an implementation of conditional inference trees within the R system for statistical computing (R Development Core Team 2004). Until the package is published on CRAN (`http://CRAN.R-project.org`) it is available upon request.

Tree Pipit Abundance.

The impact of certain environmental factors on the population density of the tree pipit *Anthus trivials* is investigated by Müller and Hothorn (2004). The occurrence of tree pipits was recorded several times at $n = 86$ stands which were established on a long environmental gradient. Among nine environmental factors, the covariate showing the largest association to the number of tree pipits is the canopy overstorey ($P = 0.003$). Two groups of stands can be distinguished: Sunny stands with less than 40% canopy overstorey ($n = 24$) show a significantly higher density of tree pipits compared to darker stands with more than 40% canopy overstorey ($n = 62$). This result is important for management decisions in forestry enterprises: Cutting the overstorey with release of old oaks creates a perfect habitat for this indicator species of near natural forest environments.

Glaucoma & Laser Scanning Images.

Laser scanning images taken from the eye background are expected to serve as the basis of an automated system for glaucoma diagnosis. Although prediction is more important in this application (Mardin, Hothorn, Peters, Jünemann, Nguyen, and Lausen 2003), a simple visualization of the regression relationship is useful for comparing the structures inherent in the learning sample with analytic expertise. For 98 patients and 98 controls, matched by age and gender, 62 covariates describing the eye morphology are available. The data is part of the `ipred` package (Peters, Hothorn, and Lausen 2002, `http://CRAN.R-project.org`). The first split in Figure 1 separates eyes with a volume above reference less than 0.059 mm^3 in the inferior part of the optic nerve head (`vari`). Observations with larger volume are mostly controls, a finding which corresponds with analytic expertise: The volume above reference measures the thickness of the nerve layer, expected to decrease with a glaucomatous damage of the optic nerve. Further separation is achieved by the

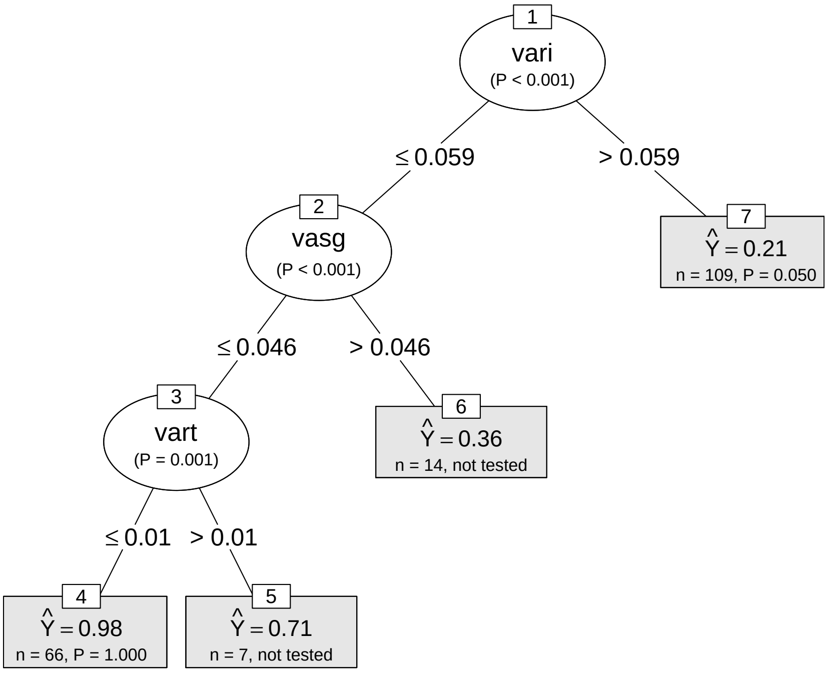

Figure 1: Binary conditional inference tree for the glaucoma data. For each node, the Bonferroni-adjusted P-values are given, the fraction of glaucomatous eyes is displayed for each terminal node.

volume above surface global (`vasg`) and the volume above reference in the temporal part of the optic nerve head (`vart`).

Node Positive Breast Cancer.

Recursive partitioning for censored responses has attracted a lot of interest (e.g., Segal 1988; LeBlanc and Crowley 1992). Survival trees using P-value adjusted Logrank statistics are used by Schumacher, Holländer, Schwarzer, and Sauerbrei (2001) for the evaluation of prognostic factors for the German Breast Cancer Study Group (GBSG2) data, a prospective, controlled clinical trial on the treatment of node positive breast cancer patients. Here, we use Logrank scores as well. Complete data of seven prognostic factors of 686 women are used for prognostic modeling, the dataset is available within the `ipred` package. The number of positive lymph nodes (`pnodes`) and the progesterone receptor (`progrec`) have been identified as prognostic factors in the survival tree analysis by Schumacher *et al.* (2001). Here, the binary variable coding whether a hormonal therapy was applied or not (`horTh`) additionally is part of the model depicted in Figure 2.

Mammography Experience.

Ordinal response variables are common in investigations where the response is a subjective human interpretation. We use an example given by Hosmer and Lemeshow (2000), p. 264, studying the relationship between the mammography experience (never, within a year, over one year) and opinions about mammography expressed in questionnaire answered by $n = 412$ women. The resulting partition based on scores $\xi = (1, 2, 3)$ is given in Figure 3. Women who (strongly) agree with the question 'You do not need a mammogram unless you develop symptoms' seldomly have experienced a mammography. The variable `benefit` is a score with low values indicating a strong agreement with the benefits of the examination. For those women in (strong) disagreement

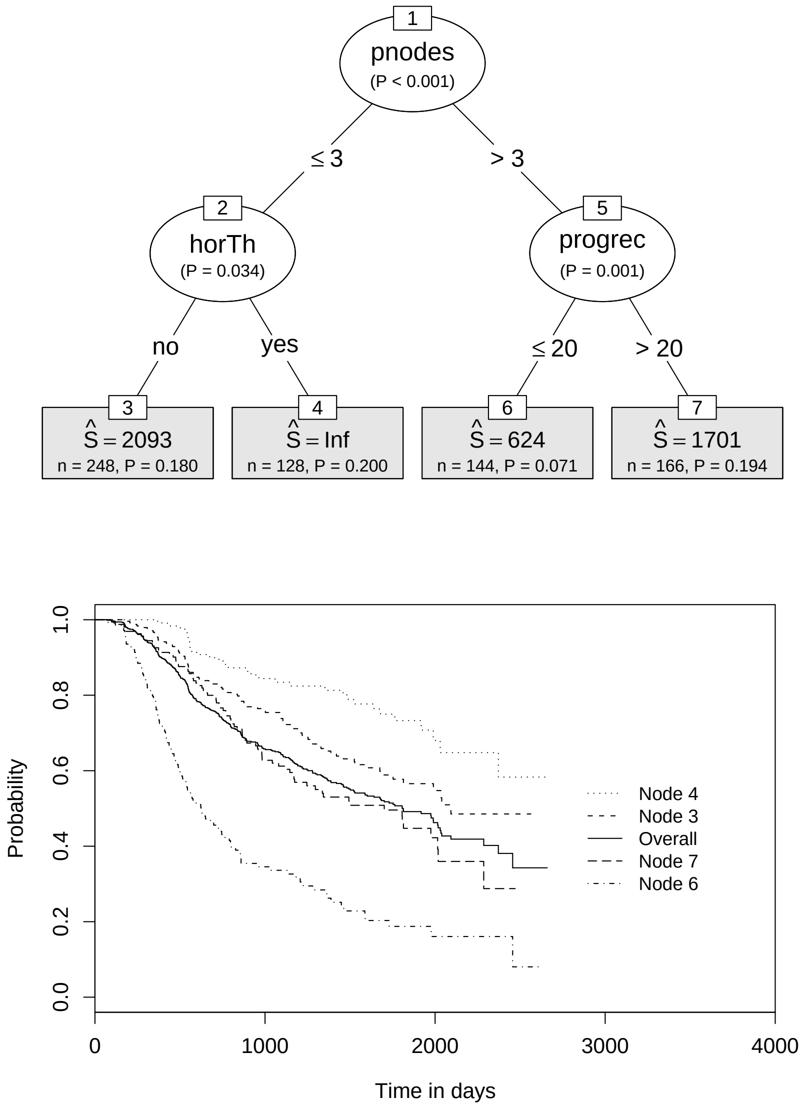

Figure 2: Tree-structured survival model for the GBSG2 data (top) and the distribution of survival times in the terminal nodes (bottom). The median survival time is displayed in each terminal node of the tree.

with the first question above, low values of `benefit` identify persons being more likely to have experienced such an examination at all.

Hunting Spiders.

Finally, we take a closer look at a challenging dataset on animal abundance first reported by Van der Aart and Smeenk-Enserink (1975) and re-analyzed by De'ath (2002) using regression trees dealing with multivariate responses. The abundance of 12 hunting spider species is to regressed on six environmental variables (`water`, `sand`, `moss`, `reft`, `twigs` and `herbs`) for $n = 28$ observations. Because of the small sample size we allow for a split if at least 5 observations are element of a node and approximate the distribution of a $c_{\max}$ test statistic by 9999 Monte-Carlo replications. The prognostic factors found by De'ath (2002), `twigs` and `water`, are confirmed by the model shown in Figure 4 which additionally identifies `reft`. The data are available in package `mvpart` at `http://CRAN.R-project.org`.

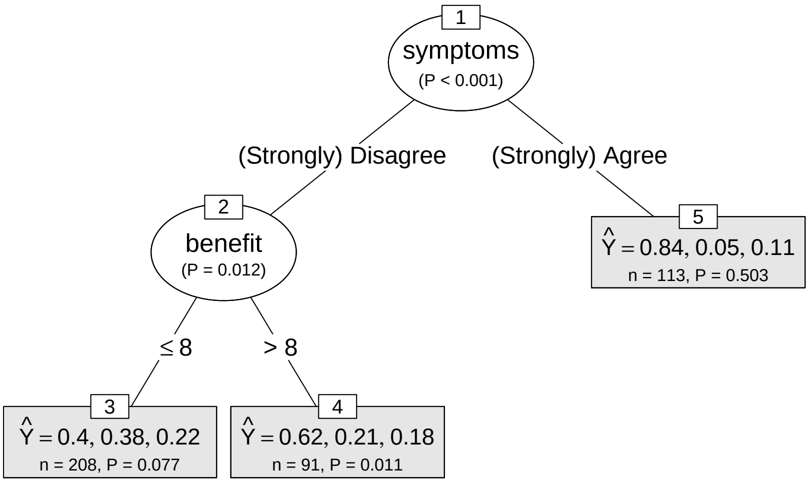

Figure 3: Ordinal regression for the mammography experience data with the fractions of (never, within a year, over one year) given in the nodes. No admissible split was found for node 4.

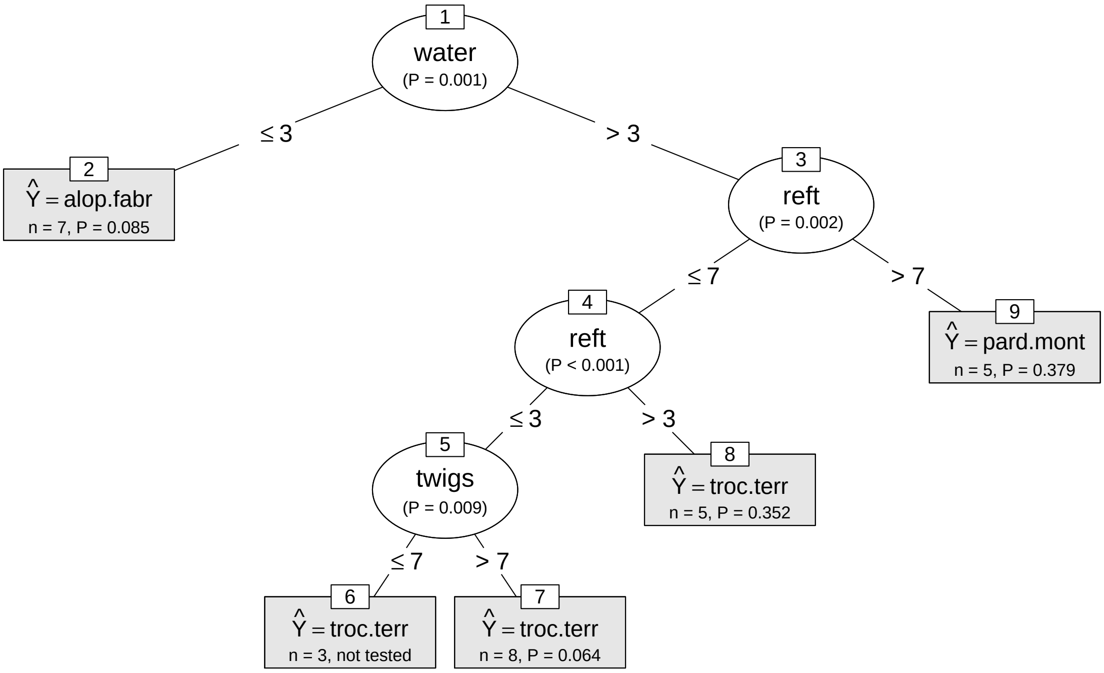

Figure 4: Regression tree for hunting spider abundance. The species with maximal abundance is given for each terminal node.

6. Empirical Comparisons

In this section, we compare the conditional inference trees developed in this paper with the well established trees obtained by exhaustive search in combination with pruning methods. With respect to explanatory performance, we show that the unbiased partitioning indeed leads to structurally different partitions of the data. Furthermore, the predicitive performance of conditional inference trees can be shown to be equivalent to that of optimally pruned trees.

The `rpart` package (Therneau and Atkinson 1997) essentially implements the algorithms described in the CART book by Breiman *et al.* (1984) and is the de-facto standard in open-source recursive partitioning software. It implements cost-complexity pruning based on cross-validation after an initial large tree was grown by an exhaustive search. The conditional inference trees are constructed with the same restrictions as given in Section 5, based on c_{quad}-type test statistics and a Bonferroni correction with $\alpha = 0.05$. The empirical experiments were performed in the R system for statistical computing (version 1.9.1 R Development Core Team 2004) using 10 benchmarking problems from the UCI repository (Blake and Merz 1998) as well as the glaucoma data. Some characteristics of the problems are given in Table 1. We draw 500 random samples from the out-of-bag performance measures (misclassification or mean-squared error) in a dependent two-sample design as described in the conceptual framework for benchmark experiments of Hothorn, Leisch, Zeileis, and Hornik (2004).

The performance distributions are said to be equivalent when the performance of the conditional inference trees (with $\alpha = 0.05$) compared to the performance of optimally pruned `rpart` trees does not differ by an amount of more than 10%. The null hypothesis of non-equivalent performances is rejected at the 5% level when the 90% two-sided Fieller confidence interval for the ratio of the expectations is completely included in the equivalence range $(0.9, 1.1)$.

We measure the discrepancy between the partitions from both methods by the normalized mutual information ('NMI' Strehl and Ghosh 2002), essentially the ratio of the mutual information of two partitions standardized by the entropy of both partitions. Values near one indicate similar to equal partitions while values near zero are obtained for structurally different partitions. The distribution of the performance ratios together with the Fieller confidence intervals are depicted in Figure 5.

Equivalent performance cannot be postulated for the Glass data: The performance of the conditional inference trees is roughly 10% worse compared with `rpart`. In addition, the null hypothesis of non-equal performance cannot be rejected for the Boston Housing problem, the Ionosphere data and the Ozone example. Here, the conditional inference trees perform better compared to `rpart` trees by a magnitude of 25% (Boston Housing), 10% (Ionosphere) and 15% (Ozone). For all other problems, the performance of optimally pruned `rpart` trees and trees fitted within a permutation testing framework can be assumed to be equivalent. However, the median normalized mutual information is 0.447 and a bivariate density estimate depicted in Figure 6 does not indicate any relationship between the ratio of the performances and the discrepancy of the partitions. Therefore, we can conclude that the partitions and thus the models induced by both algorithms are structurally different.

7. Discussion

Recursive binary partitioning, a popular tool for regression analysis, is embedded into a well defined framework of conditional inference procedures. Both the overfitting and variable selection problem induced by a recursive fitting procedure are solved by means of statistical test procedures. Nevertheless, one should keep in mind that regression models based on univariate rectangular splits can only serve as a rough approximation to reality. Conclusions drawn from tree structures visualized in a way similar to Figures 1–4 are valid in a sense that covariates without association to the response appear in a node only with a probability not exceeding α. Therefore, the conditional inference trees suggested in this paper are not just heuristics but non-parametric models with well defined theoretical background. While the predictions obtained from conditional inference trees

	J	m	n	NA	nominal	ordinal	continuous
Boston Housing	–	13	506	–	–	–	13
Breast Cancer	2	9	699	16	4	5	–
Diabetes	2	8	768	–	–	–	8
Glass	6	9	214	–	–	–	9
Glaucoma	2	62	196	–	–	–	62
Ionosphere	2	33	351	–	1	–	32
Ozone	–	12	361	158	3	–	9
Servo	–	4	167	–	4	–	–
Sonar	2	60	208	–	–	–	60
Soybean	19	35	683	121	35	5	–
Vehicle	4	19	846	–	–	–	19
Vowel	11	10	990	–	1	–	9

Table 1: Summary of the benchmarking problems showing the number of classes of a nominal response J ('–' indicates a continuous response), the number of observations with at least one missing value (NA) as well as the measurement scale of the covariates.

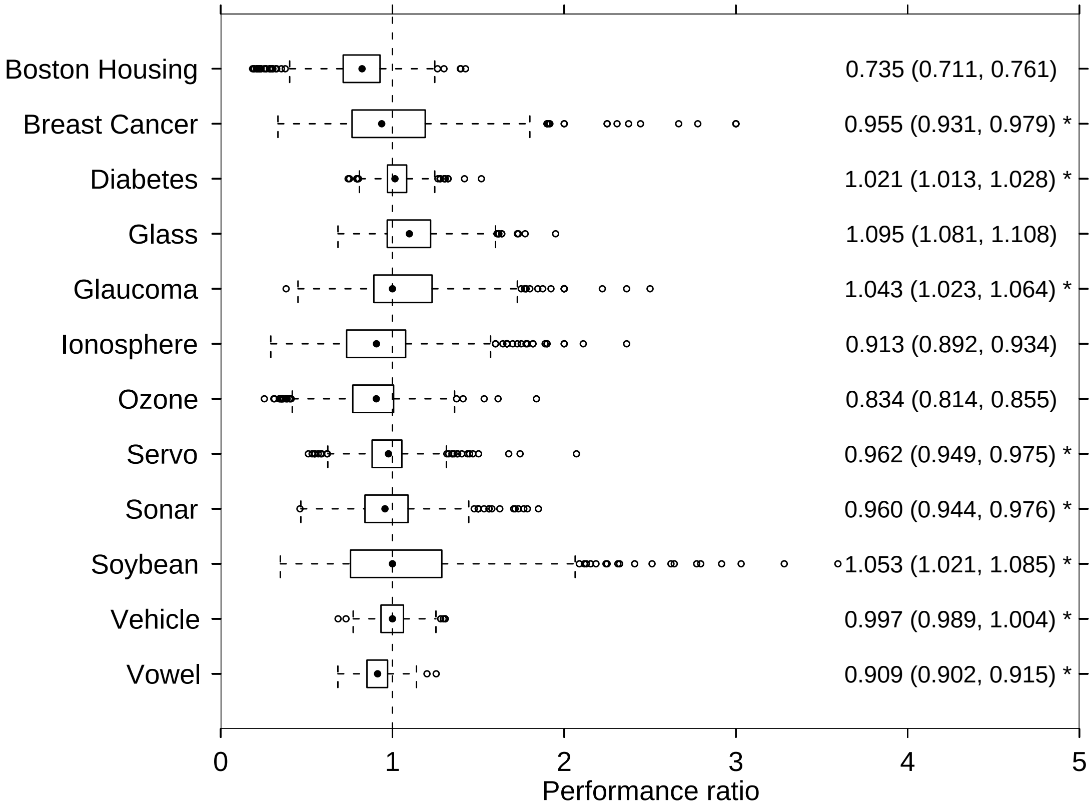

Figure 5: Distribution of the pairwise ratios of the performances of the conditional inference trees and `rpart` accomplished by estimates and 90% Fieller confidence intervals for the ratio of the expectations of the performance distributions. Stars indicate equivalent performances.

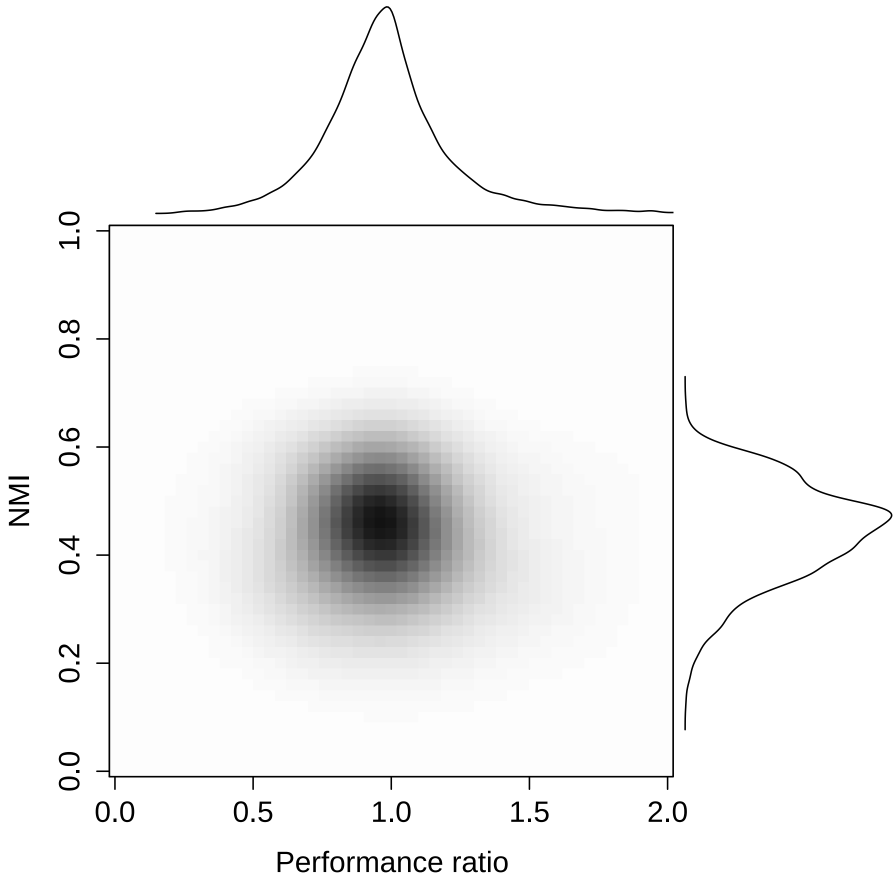

Figure 6: Distribution of the pairwise performance ratios and the normalized mutual information measuring the discrepancy of the induced partitions.

are as good as the predictions of pruned trees, the partitions induced by both algorithms differ structurally. Therefore, the interpretations obtained from conditional inference trees and trees fitted by an exhaustive search without bias correction can not be assumed to be equivalent.

In addition to its advantageous statistical properties, our framework is computationally attractive. The computational complexity of the algorithm is of order n and, for nominal covariates measured at K levels, the evaluation of all $2^{K-1} - 1$ possible splits is not necessary for the variable selection. In contrast to algorithms incorporating pruning based on resampling, the models suggested here can be fitted deterministically. Although we restricted ourself to binary splits, the incorporation of multiway splits in step 2 of the algorithm is possible, for example utilizing the work of O'Brien (2004).

References

Agresti A (2002). *Categorical Data Analysis*. John Wiley & Sons, New York, 2nd edition.

Blake C, Merz C (1998). "UCI Repository of machine learning databases." URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and regression trees.* Wadsworth, California.

De'ath G (2002). "Multivariate Regression Trees: A New Technique For Modeling Species-Environment Relationships." *Ecology*, **83**(4), 1105–1117.

Dobra A, Gehrke J (2001). "Bias Correction in Classification Tree Construction." In "Proceedings of the Eighteenth International Conference on Machine Learning," pp. 90–97. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

Frank E, Witten IH (1998). "Using a Permutation Test for Attribute Selection in Decision Trees." In "Proceedings of the Fifteenth International Conference on Machine Learning," pp. 152–160. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Genz A (1992). "Numerical computation of multivariate normal probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression.* John Wiley & Sons, New York, 2nd edition.

Hothorn T, Leisch F, Zeileis A, Hornik K (2004). "The design and analysis of benchmark experiments." *Technical Report 82*, SFB Adaptive Informations Systems and Management in Economics and Management Science. URL http://www.wu-wien.ac.at/am/reports.htm#82.

Jensen DD, Cohen PR (2000). "Multiple Comparisons in Induction Algorithms." *Machine Learning*, **38**, 309–338.

Kass G (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics*, **29**(2), 119–127.

Kim H, Loh WY (2001). "Classification Trees With Unbiased Multiway Splits." *Journal of the American Statistical Association*, **96**(454), 589–604.

Larsen DR, Speckman PL (2004). "Multivariate Regression Trees for Analysis of Abundance Data." *Biometrics*, **60**, 543–549.

Lausen B, Hothorn T, Bretz F, Schumacher M (2004). "Optimally Selected Prognostic Factors." *Biometrical Journal*, **46**(3), 364–374.

Lausen B, Schumacher M (1992). "Maximally Selected Rank Statistics." *Biometrics*, **48**, 73–85.

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

LeBlanc M, Crowley J (1993). "Survival Trees by Goodness of Split." *Journal of the American Statistical Association*, **88**(422), 457–467.

Loh WY (2002). "Regression Trees With Unbiased Variable Selection And Interaction Detection." *Statistica Sinica*, **12**, 361–386.

Loh WY, Shih YS (1997). "Split Selection Methods for Classification Trees." *Statistica Sinica*, **7**, 815–840.

Loh WY, Vanichsetakul N (1988). "Tree-Structured Classification via Generalized Discriminant Analysis." *Journal of the American Statistical Association*, **83**, 715–725. With discussion.

Mardin CY, Hothorn T, Peters A, Jünemann AG, Nguyen NX, Lausen B (2003). "New Glaucoma Classification Method based on standard HRT parameters by bagging classification trees." *Journal of Glaucoma*, **12**(4), 340–346.

Molinaro AM, Dudoit S, van der Laan MJ (2004). "Tree-Based Multivariate Regression and Density Estimation with Right-Censored Data." *Journal of Multivariate Analysis*, **90**(1), 154–177.

Morgan JN, Sonquist JA (1963). "Problems in the analysis of survey data, and a proposal." *Journal of the American Statistical Association*, **58**, 415–434.

Müller J, Hothorn T (2004). "On the identification and assessment of habitat patterns with impact on breading bird communities in oak forests." *European Journal of Forest Research*. (accepted).

Noh HG, Song MS, Park SH (2004). "An unbiased method for constructing multilabel classification trees." *Computational Statistics & Data Analysis*. (in press).

O'Brien SM (2004). "Cutpoint Selection for Categorizing a Continuous Predictor." *Biometrics*, **60**, 504–509.

Peters A, Hothorn T, Lausen B (2002). "ipred: Improved Predictors." *R News*, **2**(2), 33–36. ISSN 1609-3631, URL `http://CRAN.R-project.org/doc/Rnews/`.

Quinlan JR (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publ., San Mateo, California.

Rasch D (1995). *Mathematische Statistik*. Johann Ambrosius Barth Verlag, Heidelberg, Leipzig.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.

Rounds EM (1980). "A Combined Nonparametric Approach To Feature Selection and Binary Decision Tree Design." *Pattern Recognition*, **12**, 313–317.

Schumacher M, Holländer N, Schwarzer G, Sauerbrei W (2001). "Prognostic Factor Studies." In J Crowley (ed.), "Statistics in Oncology," pp. 321–378. Marcel Dekker, New York, Basel.

Segal MR (1988). "Regression Trees for Censored Data." *Biometrics*, **44**, 35–47.

Shih Y (2004). "A note on split selection bias in classification trees." *Computational Statistics & Data Anlysis*, **45**, 457–466.

Strasser H, Weber C (1999). "On the asymptotic theory of permutation statistics." *Mathematical Methods of Statistics*, **8**, 220–250.

Strehl A, Ghosh J (2002). "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions." *Journal of Machine Learning Research*, **3**, 583–617.

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

Therneau TM, Atkinson EJ (1997). "An Introduction to Recursive Partitioning using the rpart Routine." *Technical Report 61*, Section of Biostatistics, Mayo Clinic, Rochester. URL `http://www.mayo.edu/hsr/techrpt/61.pdf`.

Van der Aart PJ, Smeenk-Enserink N (1975). "Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environment characteristics in a dune area." *Netherlands Journal of Zoology*, **25**, 1–45.

Westfall PH, Young SS (1993). *Resampling based Multiple Testing*. John Wiley & Sons, New York.

White AP, Liu WZ (1994). "Bias in Information-based Measures in Decision Tree Induction." *Machine Learning*, **15**, 321–329.

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

Appendix

An equivalent but computational simpler formulation of the linear statistic for case weights greater one can be written as follows. Let $\mathbf{a} = (a_1, \dots, a_{\mathbf{w}\cdot})$, $a_l \in \{1, \dots, n\}, l = 1, \dots, \mathbf{w}\cdot$, denote the vector of observation indices, with index i occuring w_i times. For one permutation σ of $\{1, \dots, \mathbf{w}\cdot\}$, the linear statistic (1) may be written as

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = \operatorname{vec}\left(\sum_{k=1}^{\mathbf{w}\cdot} g_j(X_{ja_k}) h(\mathbf{Y}_{\sigma(\mathbf{a})_k}, (\mathbf{Y}_1, \dots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^{p_j q}.$$