Gaining Insight with Recursive Partitioning of Generalized Linear Models

Thomas Rusch
WU Wirtschaftsuniversität Wien

Achim Zeileis
Universität Innsbruck

Abstract

Recursive partitioning algorithms separate a feature space into a set of disjoint rectangles. Then, usually, a constant in every partition is fitted. While this is a simple and intuitive approach, it may still lack interpretability as to how a specific relationship between dependent and independent variables may look. Or it may be that a certain model is assumed or of interest and there is a number of candidate variables that may non-linearly give rise to different model parameter values. We present an approach that combines generalized linear models with recursive partitioning that offers enhanced interpretability of classical trees as well as providing an explorative way to assess a candidate variable's influence on a parametric model. This method conducts recursive partitioning of a generalized linear model by (1) fitting the model to the data set, (2) testing for parameter instability over a set of partitioning variables, (3) splitting the data set with respect to the variable associated with the highest instability. The outcome is a tree where each terminal node is associated with a generalized linear model. We will show the method's versatility and suitability to gain additional insight into the relationship of dependent and independent variables by two examples, modelling voting behaviour and a failure model for debt amortization, and compare it to alternative approaches.

Keywords: model-based recursive partitioning, generalized linear models, model trees, functional trees, parameter instability, maximum likelihood.

1. Introduction

In many fields, classic parametric models are still dominant in statistical modelling and often rightly so. They demand some insight into the data generating process as well as a strong theoretical foundation to be applicable and as such force a researcher to be clear about the question she wants answered and to put a great deal of thought into collecting data and setting up the statistical model. They have the undeniable advantage to be interpretable in light of the research questions. Usually they pose restrictions on the relationship between the explanatory variables and the target variables. A very common restriction is to define the functional relationship between (transformations of) the independent and (transformations of) the dependent variables as linear. This gives rise to many parametric models, such as the classic linear model (Rao and Toutenburg 1997), generalized linear models (GLM, McCullagh and Nelder 1989) or, somewhat more generally, maximum likelihood (ML) models with linear predictors (LeCam 1990).

However, the linearity assumption for the coefficients of the predictor variables is precisely

This is an electronic version of an article published in *Journal of Statistical Computation and Simulation*, **83**(7), 1301–1315, available online doi:10.1080/00949655.2012.658804

what can sometimes appear to be too rigid for the whole data set, even if the model might fit well in a subsample. Especially with large data sets or data sets where knowledge about the underlying processes is limited, setting up useful parametric models can be difficult and their performance may not be sufficient. This is why a number of flexible methods that only need very few assumptions have recently been developed (sometimes collected under the umbrella terms "data mining" and "machine learning", Clarke, Fokoue, and Zhang 2009). Many of these methods are able to incorporate non-linear relationships or find the functional relationship by themselves and therefore can have higher predictive power in settings where classic models are biased or even fail. However, they may leave the researcher puzzled as to what the underlying mechanisms are, since many of them are either black box methods (e.g., random forests) or have a high variance themselves (e.g., trees). See Hastie, Tibshirani, and Friedman (2009) for a comprehensive discussion of some of the most popular of these methods and their advantages and disadvantages over classic parametric models.

In this paper we present an approach that integrates classic generalized linear models and maximum likelihood models with a linear predictor with a popular data mining method, recursive partitioning or trees. Trees have become a widely researched method since their first inception by Morgan and Sonquist (1968), see e.g., Breiman, Friedman, Olshen, and Stone (1984), Quinlan (1993), Hothorn, Hornik, and Zeileis (2006), Zhang and Singer (2010). Their biggest advantage is often seen in being simple to interpret and easy to visualize and at the same time allowing to incorporate high-order interactions and exhibiting higher predictive power than classic approaches. Over the last 20 years, effort went into combining parametric regression models with recursive partitioning (Chaudhuri, Lo, Loh, and Yang 1995). These approaches were sometimes coined hybrid, model or functional trees (Gama 2004) and include methods such as M5 (Quinlan 1993), SUPPORT (Chaudhuri, Huang, Loh, and Yao 1994), GUIDE (Loh 2002), LMT (Landwehr, Hall, and Eibe 2005) and LOTUS (Chan and Loh 2004). A recent proposal is model-based recursive partitioning (MOB, Zeileis, Hothorn, and Hornik 2008) which provides a unified framework for fitting, splitting and pruning based on M-estimation (including least squares and maximum likelihood as special cases).

Building upon the MOB framework, in what follows we explicitly present and discuss recursive partitioning of generalized linear and related models. The remainder of the paper is as follows: In Section 2 we discuss recursive partitioning of generalized linear models, from the basic idea of MOB in Section 2.1 and generalized linear models in Section 2.2 to the specific algorithm in Section 2.3. In Section 2.4 we discuss the extension to models with linear predictors that do not strictly belong to the class of GLM. In Section 3 we illustrate the usage of the algorithm for two data sets and how additional insight can be gained from this hybrid approach. Section 4 contains a comparative investigation into similarities and difference in applicability, properties and performance of the presented approach with alternative approaches from the literature. We conclude with a general discussion in Section 5.

2. Recursive partitioning of generalized linear models

2.1. Basic idea

Model-based recursive partitioning (Zeileis *et al.* 2008) looks for a piece-wise (or segmented) parametric model $\mathcal{M}_{\mathcal{B}}(Y, \{\boldsymbol{\vartheta}_b\}), b = 1, \ldots, B$ that may fit the data set at hand better than a

global model $\mathcal{M}(Y, \boldsymbol{\vartheta})$, where Y are observations from a space $\mathcal{Y}$. The existence of the real p-dimensional parameter vector in each segment $\boldsymbol{\vartheta}_b \in \Theta_b$ is assumed and their collection is denoted as $\{\boldsymbol{\vartheta}_b\}$. The partition $\{\mathcal{B}_b\}, b = 1, \ldots, B$ of the space $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_l$ spanned by the l covariates $Z_j, j = 1, \ldots, l$ gives rise to B segments within the data for which local parametric models $\mathcal{M}_b(Y, \boldsymbol{\vartheta}_b), b = 1, \ldots, B$ may fit better than the global model. All these local models have the same structural form, they only differ in terms of $\boldsymbol{\vartheta}_b$. Minimizing the objective function $\sum_{b=1}^{B} \sum_{i \in I_b} \Psi(Y_i, \boldsymbol{\vartheta}_b)$ (with the corresponding indices $I_b, b = 1, \ldots, B$) over all conceivable partitions $\{\mathcal{B}_b\}$ will result in the set of vectors of parameter estimates $\{\hat{\boldsymbol{\vartheta}}_b\}$. Technically this is difficult to achieve and a greedy forward search of selecting only one covariate in each step is suggested to approximate the optimal partition. In what follows, we will focus on generalized linear models (McCullagh and Nelder 1989) as the node model $\mathcal{M}(Y, \boldsymbol{\vartheta})$ and briefly extend it to other maximum likelihood models with linear predictors.

2.2. Generalized linear models

Let $Y = (y, \mathbf{x})$ denote a set of a response y and p-dimensional covariate vector $\mathbf{x} = (x_1, \ldots, x_p)$ with expected value $E(y) = \mu$. For $i = 1, \ldots, n$ independent observations, the distribution of each y_i is an exponential family with density (Aitkin, Francis, Hinde, and Darnell 2009)

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - \gamma(\theta_i)]/\phi + \tau(y_i, \phi)\} \tag{1}$$

Here, the parameter of interest (natural or canonical parameter) is θ_i, ϕ is a scale parameter (known or seen as a nuisance) and γ and τ are known functions. The n-dimensional vectors of fixed input values for the p explanatory variables are denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_p$. We assume that the input vectors influence (1) only via a linear function, the linear predictor, $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ upon which θ_i depends. As it can be shown that $\theta = (\gamma')^{-1}(\mu)$, this dependency is established by connecting the linear predictor η and θ via the mean (Venables and Ripley 2002). More specifically, the mean μ is seen as an invertible and smooth function of the linear predictor, i.e.,

$$g(\mu) = \eta \text{ or } \mu = g^{-1}(\eta) \tag{2}$$

The function $g(\cdot)$ is called the link function. If the function connects μ and θ such that $\mu \equiv \theta$, then this link is called canonical and has the form $g = (\gamma')^{-1}$. Mean and variance for the n observations are given by

$$\mathrm{E}(y_i) = \mu_i = \gamma'(\theta_i) \qquad \mathrm{Var}(y_i) = \phi\gamma''(\theta_i) = V_i, \tag{3}$$

with $'$ and $''$ denoting the first and second derivatives respectively. Considering the GLM $\eta_i = g(\mu_i) = \boldsymbol{\beta}'\mathbf{x}_i$, the log-likelihood for n observations is given by Aitkin *et al.* (2009)

$$l(\theta, \phi; Y) = \sum_{i=1}^{n} [y_i\theta_i - \gamma(\theta_i)]/\phi + \sum_{i=1}^{n} \tau(y_i, \phi). \tag{4}$$

The score functions for $\boldsymbol{\beta}$ are then Aitkin *et al.* (2009)

$$S(\boldsymbol{\beta}, y_i) = \frac{\partial l(\boldsymbol{\beta}, \phi; Y)}{\partial \boldsymbol{\beta}} = \sum_i (y_i - \mu_i)\mathbf{x}_i / V_i g'(\mu_i), \tag{5}$$

and the information matrix is,

$$\begin{aligned} \mathcal{I}(\hat{\boldsymbol{\beta}}) &= -\frac{\partial^2 l(\boldsymbol{\beta}, \phi; Y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \\ &= -\sum_i \mathbf{x}_i \mathbf{x}_i' / V_i g_i^{'2} - \sum_i (y_i - \mu_i) \mathbf{x}_i \mathbf{x}_i' (V_i g_i'' + V_i' g_i') / V_i^2 g_i^{'3}, \end{aligned} \tag{6}$$

with $g_i' = g'(\mu_i)$, $V_i' = \frac{dV_i}{d\mu_i}$ and $g_i'' = \frac{d^2 g(\mu_i)}{d\mu_i^2}$. In classic GLM the observed and expected information matrix has a block-diagonal structure so the cross-derivatives of $\boldsymbol{\beta}$ and ϕ are zero. Also, the structure of (5) shows that the MLE for $\boldsymbol{\beta}$ can be obtained independently of the nuisance parameter.

Asymptotically, the estimated parameter vector $\hat{\boldsymbol{\beta}}$ shows the same properties as other ML estimators (McCullagh and Nelder 1989) and is

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}_{p+1}(\mathbf{0}, \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}), \tag{7}$$

under standard regularity conditions.

2.3. Recursive partitioning algorithm

For GLM as described earlier, the algorithm of Zeileis *et al.* (2008) becomes:

1. Fit a generalized linear model (2) to all observations in the current node b. Hence, $\boldsymbol{\beta}_b$ is estimated by minimizing the negative of the log-likelihood (4). This can be achieved by setting the score function (5) to zero (which is admissible under mild regularity conditions) to yield the estimated parameter vector $\hat{\boldsymbol{\beta}}_b$.

2. Assess stability of the score function evaluated at the estimated parameter, $\hat{s}_i = S(\hat{\boldsymbol{\beta}}_b, y_i)$ with respect to every possible ordering of the values of each partitioning covariates $Z_j, j = 1, \ldots, l$ with generalized M-fluctuation tests (Zeileis and Hornik 2007). This yields a measure of instability of the parameter estimates for each covariate. If there is significant instability for one or more Z_j, select the Z_j associated with the highest instability. Here the p-value of the fluctuation test is used as a measure of effect size, the lower the p-value the higher the associated instability. If no significant instability is found, the algorithm stops. Please note that the significance level for the fluctuation tests has to be corrected for multiple testing to keep the global significance level, which can be achieved by a simple Bonferroni correction (Hochberg and Tamhane 1987).

3. After a splitting variable has been selected, the split points are computed by locally optimizing $-\sum_{k=1}^{K} l(\boldsymbol{\beta}_k, \phi; y_i \mathbb{1}_{[i \in I_k]})$ with $\mathbb{1}_{[\cdot]}$ denoting the indicator function. In principle this can be done for any number $K - 1$ of fixed or adaptively chosen splits that is less or equal to the number of observations in the current node. However, we restrict ourselves to binary splits, i.e., only one split point is chosen. This means we minimize $-l(\boldsymbol{\beta}_1, \phi; y_i \mathbb{1}_{[i \in I_1]}) - l(\boldsymbol{\beta}_2, \phi; y_i \mathbb{1}_{[i \in I_2]})$ for two rival segmentations with corresponding indices I_1 and I_2 by an exhaustive search over all pairwise comparisons of possible partitions.

4. This is then repeated recursively for each daughter node until no significant instability is detected or another stopping criterion is reached.

Parameter stability tests Step 2 in the algorithm above needs some additional details. As mentioned above, the parameter stability of the individual score function contributions with respect to the splitting variable Z_j is assessed by means of generalized M-fluctuation tests (Zeileis and Hornik 2007) for any ordering of the values of $Z_j, \sigma(Z_{ij})$. For a discussion of the empirical fluctuation process of the cumulative deviations of the score function $S(\hat{\boldsymbol{\beta}}_b, y_i)$ with respect to $\sigma(Z_{ij})$, $W_j(t, \hat{\boldsymbol{\beta}})$, and its asymptotical properties we refer to Zeileis and Hornik (2007) and Zeileis (2005). Depending on the nature of the covariate, we make use of two specific M-fluctuation tests for testing the null hypothesis of parameter stability for the empirical fluctuation process, $\lambda(W_j(\cdot)) = \lambda(W^0)$ where λ is a scalar functional and W_0 is a Brownian bridge. For continuous Z_j the *supLM* statistic (Andrews 1993) is used and for categorical covariates (factors) we employ the χ^2 statistic by Hjort and Koning (2002). The *SupLM* statistics is defined as

$$\lambda_{supLM}(W_j) = \max_{i=\underline{\imath},\ldots,\bar{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n}\right)^{-1} \left|\left| W_j\left(\frac{i}{n}\right)\right|\right|_2^2, \tag{8}$$

where $[\underline{\imath}, \bar{\imath}]$ is the interval over which the potential instability point is shifted (typically defined by requiring some minimal segment size $\underline{\imath}$ and $\bar{\imath} = N - \underline{\imath}$). It is the maximization of single-shift LM statistics for all possible breakpoints in $[\underline{\imath}, \bar{\imath}]$. It has as its limiting distribution a squared, k-dimensional tied-down Bessel process (Zeileis *et al.* 2008). For categorical covariates we use

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|^{-1}}{n} \left|\left| \Delta_{I_c} W_j\left(\frac{i}{n}\right)\right|\right|_2^2, \tag{9}$$

where I_c is the set of indices of observations in category $c, c = 1, \ldots, C$ and $\Delta_{I_c} W_j$ is the increment of the empirical fluctuation process over the observations in category c. This test statistic is invariant to reordering of and within categories and captures instability for splitting data according to C categories. It has as its limiting distribution a χ^2-distribution with $df = k(C-1)$.

2.4. Beyond the GLM

One important property of standard GLM is that the parameter θ (or the parameter vector of the linear predictor) and the scale parameter ϕ are orthogonal (McCullagh and Nelder 1989). Estimates of parameters of the linear predictor $\hat{\boldsymbol{\beta}}$ are therefore (almost) independent of estimates of $\hat{\phi}$ under suitable limiting conditions (White 1982). Additionally, GLM assume that the explanatory variables do not affect the scale parameter ϕ at all (Aitkin *et al.* 2009). However, it is possible to extend the methodology used here beyond the standard GLM to incorporate (i) other distributions with non-orthogonal parameters such as the exponential distribution, the Weibull distribution or the extreme value distribution, or mixtures of exponential families such as the negative binomial distribution with unknown dispersion parameter and (ii) to use a linear predictor for the scale parameter for which parameter stability can also be assessed. In both cases, the node model $\mathcal{M}(Y, \boldsymbol{\vartheta})$ and the score functions will change. This has an effect on the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, since we need to consider that we may deal with nuisance parameter estimation as well. See e.g., Aitkin *et al.* (2009) for inference with nuisance parameters. Apart from that however, the algorithm above still applies exactly the same way as long as an M-estimation approach (Huber 2009) such as maximum likelihood

is used for parameter estimation. This is because model fitting and the parameter stability tests and hence the algorithm employ M-estimation and the according asymptotics.

3. Gaining insight

3.1. Improved explanation with additional information

Due to its explorative character, model-based recursive partitioning can reveal patterns hidden within data modelled with GLM or provide further explanation of surprising or counter-intuitive results by incorporating additional information from other covariates. The tree-like structure allows the effects of these covariates to be non-linear and highly interactive as opposed to assuming a linear influence on the linked mean.

To illustrate, we use a data set from the 2004 general election in Ohio, USA. It was the presidential election of George W. Bush vs. John F. Kerry which took place on November 2nd, 2004 and saw Bush emerging as the winner with 34 more electoral seats than his adversary. Our sample consists of 19634 people from Ohio. We have aggregate voting records of each person, such as the overall number of times a person voted as well as the number of elections she was eligible to vote. Additionally, the data set includes a number of demographic, behavioural and institutional variables, such as each voter's age, gender, the party composition of the household ("partyMix"), the voter's rank ("householdRank", here the lower the number the higher the rank) and position in the household ("householdHead"), among others. We are interested in modelling the turnout of the 2004 general election on an individual level, i.e., has the person voted or not ("gen04").

In campaigning theory and voter targeting (e.g., Malchow 2008), past voting behaviour of a person is considered to be the strongest predictor of future voting behaviour. It is usually assumed that the more often a person went voting in the past, the more likely she is to do so in the upcoming election. Statistically this is a logistic regression problem with a binary dependent variable and therefore fits into the GLM framework. The number of attended elections is used as the predictor variable. It is important to note though, that the raw count of attended elections may be misleading because a higher count does not need to be the result of a person's general disposition to be more likely to vote. We therefore use the percentage of attended elections out of all elections a person was eligible to take part in to correct for possible bias. Figure 1 shows a spine plot of the data. It can be seen that the relationship is not monotonic but appears to be quadratic. This is not in accordance with intuition or the literature on voter targeting. One would expect a higher likelihood to vote for those who have a higher percentage of attended elections.

We fitted a global logistic regression model $\mathcal{M}(Y, \boldsymbol{\beta})$ with a quadratic effect of the predictor variable,

$$g(\mu) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{10}$$

where x is the percentage of attended elections ("percentAttended"). The estimated model parameters and goodness-of-fit values of the global model are displayed in Table 1. Interpolations of the predicted values were added to the spine plot in Figure 1. The initial observation could be confirmed by the model, the quadratic term turns out to be significant.

But why would people with a very high general attendance rate have a similarly low attendance

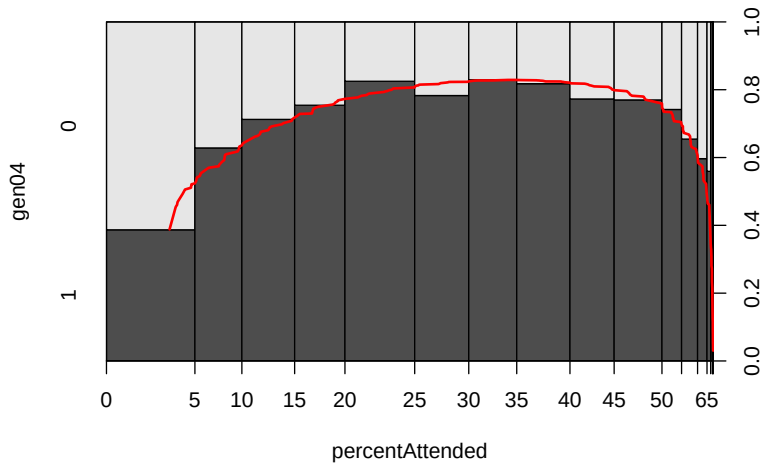

Figure 1: Spine plot of relative voting frequencies against the percent of attended elections out of all elections a person was eligible to. The solid black line is the interpolated prediction from a logistic regression model with a quadratic term for the predictor "percentAttended".

rate in the 2004 election as people who usually will attend elections rarely? And what people are they? We employ recursive partitioning of the logistic regression model in (10) to see if additional variables can shed more light on this phenomenon. We use a significance level of $\alpha = 0.05$ for the generalized M-fluctuation tests and force the minimum number of observations within each node to be at least 1060 (a fraction of about 8% of the overall data). The resulting tree is depicted in Figure 2 and the parameter estimates of the local models for the terminal nodes are given in Table 1.

Model	Node	$\hat{\beta}_0$ (se)	$\hat{\beta}_1$ (se)	$\hat{\beta}_2$ (se)	n	Dev	AIC
Global	-	−0.46 (0.03)	11.87 (0.29)	−17.32 (0.48)	19634	21948	21954
Segmented	2	$-\infty$ (−.−)	0.00 (−.−)	0.00 (−.−)	2180	0	6
	5	2.56 (0.38)	0.21 (1.87)	−6.53 (2.19)	2358	2126	2132
	7	0.42 (0.47)	14.09 (2.56)	−21.63 (3.22)	1277	808	814
	8	1.05 (0.41)	9.06 (2.17)	−15.36 (2.69)	1610	1170	1176
	10	−0.32 (0.08)	7.59 (1.17)	−4.16 (3.01)	1638	1991	1997
	13	−0.70 (0.06)	15.19 (0.77)	−19.10 (1.91)	4267	4602	4608
	14	0.16 (0.09)	12.23 (1.23)	−14.10 (3.04)	2222	1970	1976
	15	0.06 (0.14)	16.98 (1.35)	−17.82 (2.54)	4082	1565	1571

Table 1: Parameter estimates (standard errors in brackets) and goodness-of-fit statistics for the global logistic regression model and the terminal nodes of the piece-wise logistic regression model for the Ohio voter data. For legibility, $\hat{\beta}_1$ and $\hat{\beta}_2$ are given in units of the relative frequency. Please note that there are only non-voters in segment 2.

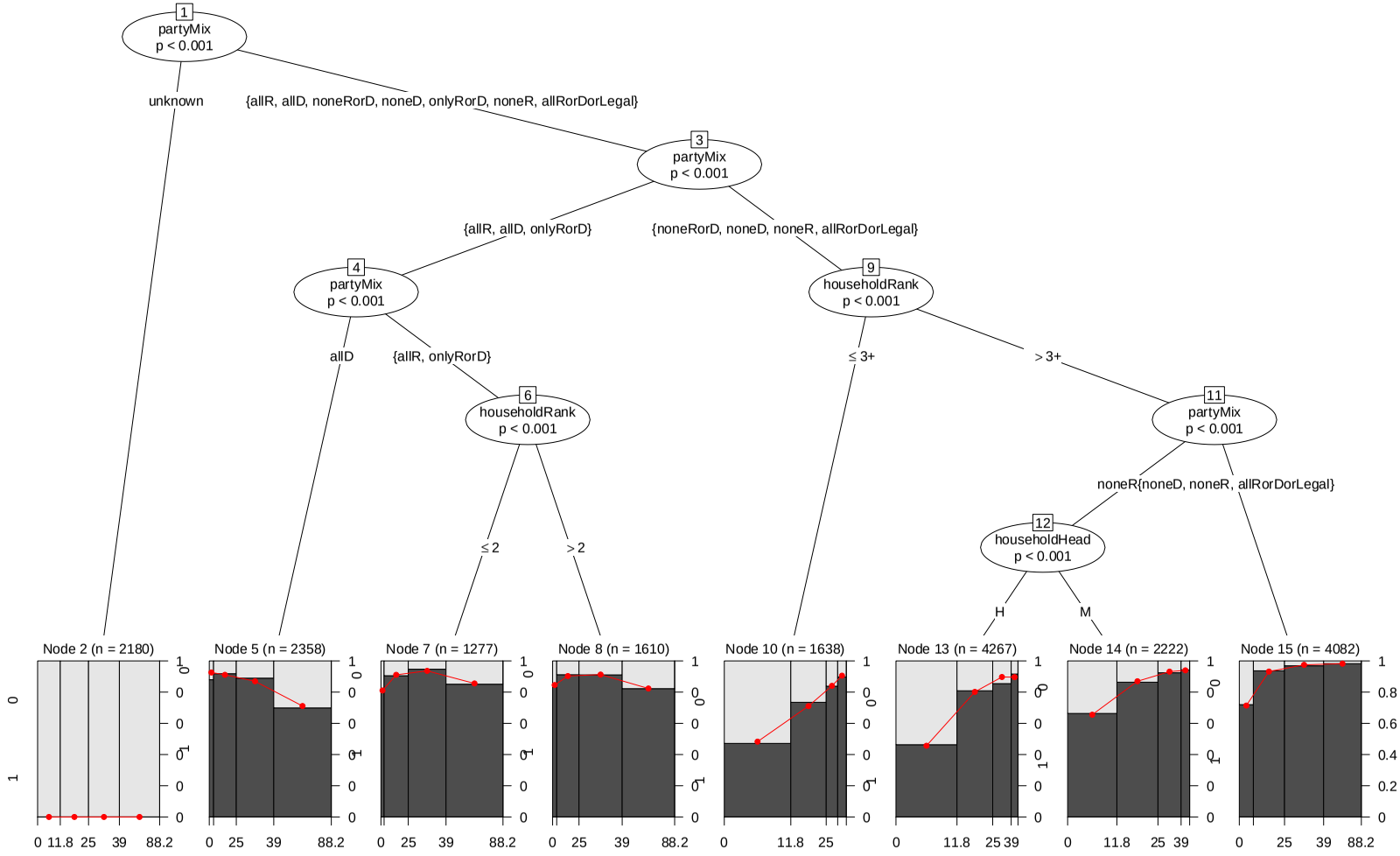

Figure 2: The resulting tree structure after partitioning the logistic regression model with linear predictor $\beta_0 + \beta_1 x + \beta_2 x^2$ where x is denoting the relative frequency of attended elections, "percentAttended". The terminal nodes display spine plots of the observed relative frequencies against the attended percentage for each partition with the solid lines connecting the predicted values from the logistic regression model.

The result from the partitioning algorithm ($\alpha = 0.05$ for the fluctuation tests) shows what or who may be responsible for the quadratic relationship between the percent of attended elections and the probability to vote. First there is a terminal node with people who did not vote at all. Please note that within this node we find (quasi-)complete separation[1](Albert and Anderson 1984). Second, the relationship is driven by the 5245 people whose household consists of members who are affiliated solely with the Democratic Party (node 5) and to a lesser extent by those affiliated solely with the Republican party (node 7) or whose household consists only of democrats and republicans (node 8). In other words, there are no independent voters in these households. Especially the segment of people whose household is composed entirely of Democrats ($n_5 = 2358$) contribute to the overall quadratic relationship seen in Figure 1. They show declining voting probability for people with a high general individual turnout and quite strongly so. While those people with a small to medium percentage of general attendance have fairly high voting probabilities that slightly increase for higher predictor values, those with a general attendance rate of 0.39 or more (nearly half of the segment) experience a sheer drop of voting probability.

For the other two segments, those whose household consists entirely of Republicans or of a mix of Republicans and Democrats ($n_7 = 1277$ and $n_8 = 1610$) this picture is less striking. Here, an attendance rate of about 0.1 to 0.4 is associated with the highest voting probability, whereas very rare voters ($x \leq 0.1$) and frequent voters ($x \geq 0.4$) have a similarly high voting probability that is slightly less than for the other people in the segment. Nodes 7 and 8 differ in the assigned rank in the household. The difference between these two nodes lies in the slightly higher overall voting probability and a higher probability for those with an attendance percentage between 25% and 40% for those with household ranks 1 and 2 (node 7).

On the other hand, the segments in terminal nodes $10, 13, 14$ and 15 indeed show a monotonically increasing voting probability for an increase of the predictor variable. This is in accordance with intuition and literature on political campaigning. Here, having at least one household member who identifies herself as "independent" is the key difference to the segments with an inverse U-shaped voting probability relationship with the percentage of attended elections. By using model-based recursive partitioning with additional covariate information, we are able to find an explanation as to why a quadratic effect has to be included into the logistic regression model. We can single out the observations that are responsible for this phenomenon and show that there are segments in which the assumed monotonic relationship is actually present.

3.2. Identifying segments with poor or good fit

Another area in which model-based recursive partitioning can be helpful is in identifying segments of the data for whom an *a priori* assumed model fits well. It may be that overall this model has a poor fit but that this is due to some contamination (for example merging two separate data files or systematic errors during data collection at a certain date). By using the described algorithm the data set might be partitioned in a way that enables us to find the segments that have poor fit and find segments for which the fit may be rather good (see also Juutilainena, Koskimäkia, Laurinena, and Röninga 2011 for an alternative for regression analysis).

[1]In this node the ML estimator does not exist. The algorithm has the positive effect of separating these observations from the rest, hence estimation in other nodes works well which might otherwise not be the case.

To illustrate this, we use data of debt amortization rates as a function of the duration of the enforcement. It can be expected that the longer the enforcement lasts, the higher amortization rate should be achieved. What is special about these data is that they came from two sources and were merged into a single data set. The merged data set consisted of $n = 165$ observations, with 75 observations from file "0" and 90 observations from file "1".

The structure of the statistical problem here is similar to a "time-to-event" analysis. We consider the amortization rate relative to the original claim as the metric variable whose hazard function we want to model. Failure to pay more, default, insolvency, bankruptcy or meeting the obligation are considered as the event "stopped paying". Additionally, we have the possibility of right censored observations if a person was lost to follow up. This lead us to using a Weibull Regression model which is an example of the type of models described in Section 2.4. Here, the scale parameter and the parameters of the linear predictor are not orthogonal and have to be estimated simultaneously.

Formally, following Venables and Ripley (2002), we model the hazard function $h(r)$, with r denoting a realization of the random variable of achieved amortization rate, R, which takes the form of

$$h(r) = \lambda^{\alpha}\alpha r^{\alpha-1} = \alpha r^{\alpha-1}\exp(\alpha\boldsymbol{\beta}^{\top}\boldsymbol{x}) \tag{11}$$

for the Weibull distribution. The parameter λ is modelled as an exponential function of the

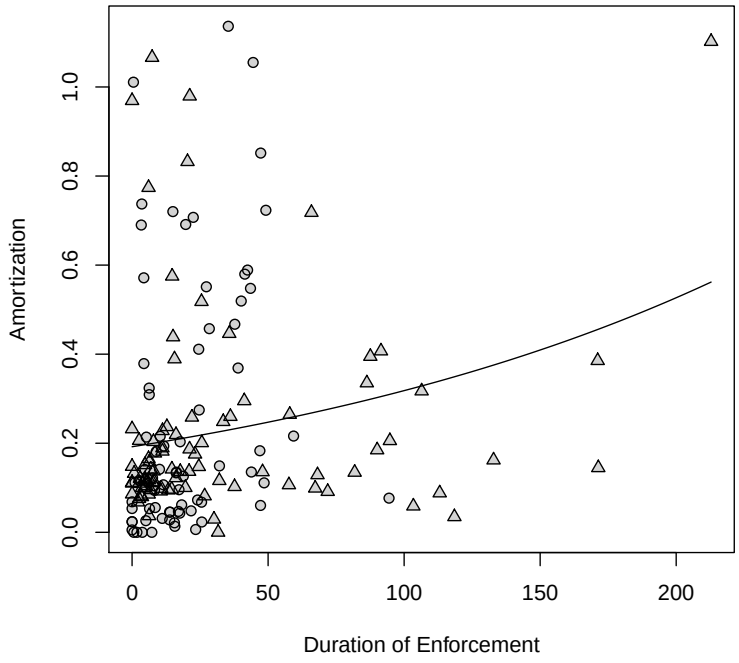

Figure 3: Scatterplot of duration of the enforcement and the achieved amortization rate until the event "failure to pay more" happened. The solid line represents the predicted values from the global Weibull regression model. Observations from file "0" are plotted as circles, those from file "1" as triangles.

Model	Node	$\hat{\beta}_0$ (se)	$\hat{\beta}_1$ (se)	Scale (sd)	n	log-lik
Global	-	−1.65 (0.12)	0.01 (0.00)	0.17 (0.06)	165	76.9
Segmented	3	−2.31 (0.33)	0.03 (0.01)	0.48 (0.11)	65	42.4
	4	1.95 (0.09)	0.01 (0.00)	−0.48 (0.09)	79	76.8
	5	−0.53 (0.22)	−0.00 (0.01)	−0.34 (0.19)	21	−4.6

Table 2: Parameter estimates (standard errors in brackets) and goodness-of-fit statistics for the global Weibull model and the terminal nodes of the segmented Weibull model for the debt amortization data.

covariates $\mathbf{x}$. In a loglinear model formulation this becomes

$$\log(R) = -\log\lambda + \frac{1}{\alpha}\log\epsilon \tag{12}$$

with ϵ being a disturbance term that is independent of $\boldsymbol{x}$ and w.l.o.g. exponentially distributed. In this particular example, $\boldsymbol{x}$ consists of an intercept and the duration of the enforcement.

A visualization of the data can be found in Figure 3. The point type corresponds to the different files, a circle for file "0" and the triangle for file "1". Additionally we include the predicted values from the global Weibull regression model. The results of the model fitting can be seen in Table 2.

What we can see here is that the model does not fit well. The log-likelihood for the regression model is 76.9 and for the intercept only model it is 75.1 which is not a significant difference at $\alpha = 0.05$ ($p = 0.054$). Apart from that it looks as though the Weibull regression is not really appropriate for the whole data set. However, one can see that a subset of the data may be appropriately modelled with the proposed relationship if it were not for the observations that have quite high amortization rates for a low enforcement duration. There are at least two possible explanations for such a lack of fit: (i) explanatory variables that were not considered in the model (misspecification) and (ii) data contamination. In this analysis it is quite likely that (i) has some effect. Hence we use information from other covariates in the subsequent recursive partitioning and gauge their influence. Inspection of Figure 3 however reveals something else. Observations that have high amortization rates for low duration time are mainly from file "0". Additionally the distribution of the enforcement duration in file "1" is more skewed (skewness 1.96 vs. 1.39) and has a much longer right tail. The same holds for the amortization rate. It looks as if merging of the two data sets could have led to a contamination as they are probably not comparable. We partition these based on the Weibull regression model from (11). As additional covariates that are used for partitioning we have the person's gender, liability at the begin of the enforcement, the current liability, the number of securities a person has as well as a person's collateralization ratio. We also include a dummy variable to flag which file the observation was from. The significance level for the parameter test is again 0.05.

The resulting tree can be found in Figure 4 and the estimated model in Table 2. We see that both suspicions from above can be confirmed. First, there is an additional variable, collateralization ratio, that seems to be relevant. Its inclusion leads to a segment where the influence of the duration is not significant. This is partly due to the small sample size in this node, but we can also see that the regression coefficient has a negative sign. It does

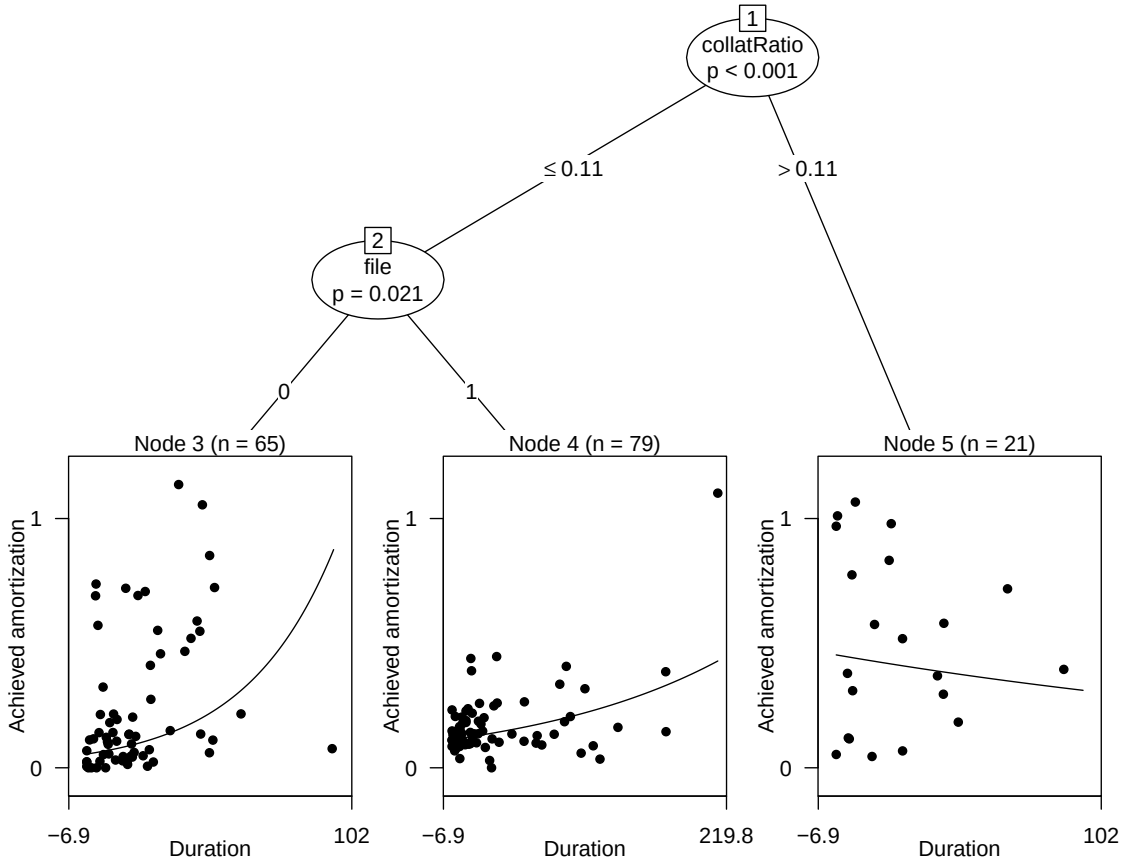

Figure 4: The recursively partitioned Weibull regression model of amortization rate explained by the duration of the enforcement. For each terminal node there is a scatterplot with the solid line representing the predicted values from the local model.

not appear as if there would be a positive relationship that we just do not detect but rather that there is no positive relationship at all. This makes sense, as the collateralization ratio is a measure of how many and how well diversified the securities of a person are and how high their value is. A person with a high collateralization ratio (two cars for example) may be able to amortize her debt very fast or at least it may not depend on the duration of the enforcement. It seems rather likely that a person with a high collateralization ratio who does not amortize her debt rather soon may have problems with or may refuse payment regardless of enforcement duration.

Second, for those with a collateralization ratio of less than 0.11, the algorithm points to a difference in the two data sources. For one data set, file "1", the Weibull regression actually fits rather well (node 4, log-likelihood of 76.8). Additionally, we have a significant positive influence of the explanatory variable. Please note however, that the coefficient and corresponding p-value is highly influenced by an outlier with amortization rate greater than 1. Removing this value leads to a much weaker association that is barely significant on a 5% level[2]. In

[2]If a semi-parametric Cox model is fitted, there is no significant influence.

node 3, for which all observations stem from file "0", we see an ill fit of the Weibull model with a log-likelihood of 45.1. It even looks as if the (significant) regression line is splitting the data in this node into two groups rather than explaining them. There seems to be heterogeneity in the data in this segment that cannot be explained by the regression model.

What we can see from this analysis however is that recursive partitioning of models can help us identify segments in our data for which the model may either fit well or may be inappropriate. Here, merging the data from file "0" with those in file "1" leads to some contamination of the merged data set. This contamination masks the acceptable fit for the subset of observations from file "1", a fact that is not necessarily clear from the non-segmented analysis. Most probably those two data sets were obtained individually and on different occasions or for different studies. They just happen to have similar variables in them. This goes to show once again that planning a study involves more than just collecting data.

4. Comparison to similar approaches

A number of model tree algorithms have been proposed in recent years. Table 3 gives an overview of different model tree algorithms, properties of the tree induction, which node models they can fit and the available software (R, R Development Core Team 2011, Weka, Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten 2009, or author binaries) to fit them.

In the machine learning literature, tree algorithms with models in each node have been around at least since the M5 algorithm of Quinlan (1993) (see also Wang and Witten 1997 for the "rational reconstruction" M5') for linear models in nodes. Another algorithm is LMT (Landwehr *et al.* 2005), which allows trees with a boosted logistic node model for binary or multinomial responses. Gama (2004) proposed an abstract framework coined "functional trees", for building tree algorithms with univariate or multivariate splits and node models.

Several model tree algorithms were also suggested in the statistical literature, for example SUPPORT (Chaudhuri *et al.* 1994, 1995), which originally suggested smoothed or unsmoothed piece-wise polynomial models, and was subsequently extended to GLM-type and survival models (Ahn and Chen 1997; Choi, Ahn, and Chen 2005; Ahn and Loh 1994; Ahn 1994b,a, 1996b,a). Many of those model tree algorithms encompassed two novel ideas: (i) unbiasedness in the split variable selection, and (ii) separation of modelling and splitting variables. Two prominent examples are GUIDE (Loh 2002, 2009) and LOTUS (Chan and Loh 2004). While the former provides capabilities for fitting models to Gaussian responses, quantile regression (Chaudhuri and Loh 2002), Poisson models (Loh 2008), proportional hazard models and longitudinal models, the latter uses similar ideas for modelling binary responses. With the MLRT algorithm (Su, Wang, and Fan 2004), some effort also went into embedding regression trees into a rigorous statistical framework based on the likelihood as an objective function which can easily be extended to model trees (a similar idea for a very specific context has been proposed by Ichinokawa and Brodziak (2010)).

Conceptually, the MOB algorithm used in the present paper belongs to the statistically motivated algorithms and combines most advantages of the aforementioned algorithms. Like GUIDE or LOTUS, it uses unbiased split selection and allows for separation of node model and splitting variables. Similar to MLRT, it uses a rigorous statistical framework of employing the same objective function to induce the tree structure and fit the node models. Compa-

rable to SUPPORT, MOB pre-prunes the trees. Furthermore, MOB provides functionality for many different type of node models that even exceeds the versatility of SUPPORT and GUIDE. Analyzed within the "functional tree" framework, MOB with more than a single explanatory variable in the node model or with interactions as splitting variables can be seen as employing multivariate splits that allow for oblique partitioning which according to Gama (2004) is an advantage especially for large data sets. Moreover, the MOB algorithm can straightforwardly be extended to feature variable selection in the node model, post-pruning of the tree or smoothing of the piece-wise function (e.g., with Chandler and Johnson 2012).

4.1. Voting data revisited

To compare the performance of the presented approach to other algorithms, we reanalyze the Ohio voter data set 3.1 with the LMT and the LOTUS.

We fit LMT with Weka (Hall *et al.* 2009) for which we employ the RWeka interface (Hornik, Buchta, and Zeileis 2009). The trees are restricted to only allow for binary splits. As LMT does not allow for separation between variables employed for the node model and for splitting respectively, all prediction variables (including the square of "percentAttended") are supplied to the algorithm. This leads to a single root note (without any splits) and hence a global logistic model with 33 parameters. For the same data, MOB uses a tree with 7 splits and 3 parameters in the node model. LMT selects all those variables that are selected by the MOB plus some additional variables, leading to the large global logistic model. The overall classification accuracy of LMT for the training sample is 0.843 whereas the MOB has a classification accuracy of 0.840. Hence, the LMT is less parsimonious (33 vs. $24 = 3 \times 8$ parameters) while the predictive accuracy on the learning sample is only slightly higher (0.843 vs. 0.840). Additionally, the quadratic relationship between attendance percentage and voting probability is not as easily intelligible as compared to the MOB.

With the LOTUS binary, we fit a model with an analogous setup for node model and splitting variables compared to the MOB as specified in Section 3.1. A maximum number of 1060 observations per node is specified and we opt for no variable selection for the node model. Everything else is set to the LOTUS default values. The resulting pruned tree (0-SE) has 12 splits and each node model has 3 estimated parameters. Hence MOB fits a more parsimonous model tree ($3 \times 13 = 39$ parameters for LOTUS vs. $3 \times 8 = 24$ parameters for MOB). At the same time MOB achieves higher classification accuracy on the training sample (0.84 vs. 0.76). As is the case with LMT, splitting variables selected by LOTUS partly coincide with those selected by MOB. On the one hand, both algorithms select "householdRank" and "partyMix" quite often for splitting (MOB five times, LOTUS five times and high up in the tree hierarchy). On the other hand, the variables "dontPhone", "compOwner", "income" and "educationLevel" are chosen for splitting only by the LOTUS (and deeper down the tree hierarchy). The biggest difference of the LOTUS to the MOB tree is that the first split is due to observations labeled "unknown" and "noneRorD" for "partyMix". This leads to a left subtree with 5 additional leaves for the LOTUS. MOB selects the same variable but only splits off observations that have a value of "unknown", which are not partitioned further. To a depth of 3, the right subtrees after the first split for MOB and LOTUS are more or less similar in terms of splitting variables and split points and therefore explanation of the quadratic relationship is comparable for both methods.

Thus, all algorithms achieve a more or less similar classification accuracy. They all agree on

		Model tree algorithm							
		FT	GUIDE	LMT	LOTUS	M5'	MLRT	MOB	SUPPORT
Tree structure	Pre-pruning		*		*			×	×
	Post-pruning	×	×	×	×	×	×	*	*
	Unbiased		×		×			×	
	Covariate type	all	all	all	all	all	all	all	metric
	Multiway splits	×		×				*	
Separate node model and splitting variables			×		×			×	
Adaptive node model		*	×	×	×	×	*	*	*
Type of node model	Gaussian	×	×			×	×	×	×
	Binomial (Logit)	*		×	×		*	×	×
	Binomial (other links)	*					*	×	
	Quasi-Binomial (Logit)	*						×	×
	Quasi-Binomial (other links)	*						×	
	Poisson	*	×				*	×	×
	Quasi-Poisson	*						×	×
	Gamma	*					*	×	×
	Inverse Gaussian	*					*	×	
	Negative Binomial	*					*	×	
	Beta	*					*	×	
	Multinomial	*		×			*	*	
	Parametric Survival	*	×				*	×	×
	Longitudinal Gaussian	*	×				*	*	×
	General Maximum Likelihood	*					*	*	
	General Quasi-Likelihood	*						*	
	Robust (M-type)	*						*	
	Quantile	*	×						
Software			author	Weka	author	Weka		R	

Table 3: Comparison of properties and applicability of different model tree algorithms. For the rows a × denotes if there already exists an implementation and ∗ denotes if an implementation is possible within the provided framework of the specific algorithm without changes (please note that the FT algorithm is on a more abstract level than all the other specific algorithms). The last row lists the availability of software packages ("author" means binaries are publicly available from the author).

"percentAttended" and its square, "partyMix" and "householdRank" to be important variables. LMT chooses a large global regression model with a high predictive accuracy. MOB and LOTUS use a much simpler logistic model, but can achieve comparable accuracy to LMT through splits (especially MOB). For this, MOB needs a lower number of splits than LOTUS which makes the MOB results easier to interpret.

5. Conclusion

In this paper, we introduced recursive partitioning of generalized linear and related models as a special case of model-based recursive partitioning. We tried to illustrate how the algorithmic approach may lead to additional insight for a *a priori* assumed parametric model, especially if the underlying mechanisms are too complex to be captured by the GLM. As such, model-based recursive partitioning can automatically detect interactions, non-linearity, model misspecification, unregarded covariate influence and so on. As an exploratory tool, it can be used for complex and large data sets for which it has a number of advantages. On the one hand, compared to a global GLM, a MOB model tree can alleviate the problem of bias and model misspecification and provide a better fit. On the other hand, compared to tree algorithms with constants, the specification of a parametric model in the terminal nodes can add extra stability and therefore reduce the variance of the tree methods. Being a hybrid of trees and classic GLM-type models, the performance of MOB models usually lies between those two poles: They tend to exhibit higher predictive power than classic models but less than non-parametric trees (Zeileis *et al.* 2008). They add some complexity compared to classical model because of the splitting process but are usually more parsimonous than non-parametric trees. They show a slightly higher variance than a global model in bootstrap experiments, but much less than non-parametric trees (even pruned ones). Compared to other model tree algorithms, MOB often exhibits comparable predictive accuracy while at the same time being more parsimonous than direct competitors. Results from MOB trees are often easy to communicate and visualize. Additionally, MOB is currently the most versatile model tree algorithm and can be rigorously justified from a statistical point of view. We believe that the exploratory use of recursive partitioning of GLM-type models, particularly with the presented approach, is fruitful for researchers dealing with models with linear predictors to detect possible hidden structure and get a better grasp of what is really happening in the data at hand, especially if modelling with classical statistical methods reaches its limitations.

References

Ahn H (1994a). "Tree-Structured Exponential Regression Modeling." *Biometrical Journal*, **36**, 43–61.

Ahn H (1994b). "Tree-Structured Extreme Value Model Regression." *Communications in Statistics – Theory and Methods*, **23**, 153–174.

Ahn H (1996a). "Log-Gamma Regression Modeling through Regression Trees." *Communications in Statistics – Theory and Methods*, **25**, 295–311.

Ahn H (1996b). "Log-Normal Regression Modeling through Recursive Partitioning." *Computational Statistics & Data Analysis*, **21**, 381–398.

Ahn H, Chen J (1997). "Tree-Structured Logistic Model for Over-Dispersed Binomial Data with Application to Modeling Developmental Effects." *Biometrics*, **53**, 435–455.

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

Aitkin M, Francis B, Hinde J, Darnell R (2009). *Statistical Modelling in R*. Oxford University Press, New York.

Albert A, Anderson JA (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika*, **71**, 1–10.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, **61**, 821–856.

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

Chan KY, Loh WY (2004). "LOTUS – An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees." *Journal of Computational and Graphical Statistics*, **13**, 826–852.

Chandler G, Johnson L (2012). "Automatic Locally Adaptive Smoothing for Tree-Based Set Estimation." *Journal of Statistical Computation and Simulation*. doi:10.1080/00949655.2011.613395.

Chaudhuri P, Huang MC, Loh WY, Yao R (1994). "Piecewise-Polynomial Regression Trees." *Statistica Sinica*, **4**, 143–167.

Chaudhuri P, Lo WD, Loh WY, Yang CC (1995). "Generalized Regression Trees." *Statistica Sinica*, **5**, 641–666.

Chaudhuri P, Loh WY (2002). "Nonparametric Estimation of Conditional Quantiles Using Quantile Regression Trees." *Bernoulli*, **8**, 561–576.

Choi Y, Ahn H, Chen JJ (2005). "Regression Trees for Analysis of Count Data with Extra Poisson Variation." *Computational Statistics & Data Analysis*, **49**, 893–915.

Clarke B, Fokoue E, Zhang HH (2009). *Principles and Theory of Data Mining and Machine Learning*. Springer-Verlag, New York.

Gama J (2004). "Functional Trees." *Machine Learning*, **55**, 219–250.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009). "The Weka Data Mining Software: An Update." *SIGKDD Explorations*, **11**(1).

Hastie T, Tibshirani R, Friedman J (2009). *Elements of Statistical Learning*. 2nd edition. Springer-Verlag, New York.

Hjort NL, Koning A (2002). "Tests for Constancy of Model Parameters over Time." *Nonparametric Statistics*, **14**, 113–132.

Hochberg Y, Tamhane AC (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Hornik K, Buchta C, Zeileis A (2009). "Open-Source Machine Learning: R Meets Weka." *Computational Statistics*, **24**(2), 225–232.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Huber P (2009). *Robust Statistics*. 2nd edition. John Wiley & Sons, Hoboken.

Ichinokawa M, Brodziak J (2010). "Using Adaptive Area Stratifiaction to Standardize Catch Rates with Application to North Pacific Swordfish (Xiphias Gladius)." *Fish Res*, **106**, 249–260.

Juutilainena I, Koskimäkia H, Laurinena P, Röninga J (2011). "BUSDM – An Algorithm for the Bottom-Up Search of Departures from a Model." *Journal of Statistical Computation and Simulation*, **81**, 561–578.

Landwehr N, Hall M, Eibe F (2005). "Logistic Model Trees." *Machine Learning*, **59**, 161–205.

LeCam L (1990). "Maximum Likelihood – An Introduction." *ISI Review*, **58**, 153–171.

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

Loh WY (2008). "Regression by Parts: Fitting Visually Interpretable Models with GUIDE." In CH Chen, W Härdle, A Unwin (eds.), *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pp. 447–469. Springer-Verlag, New York.

Loh WY (2009). "Improving the Precision of Classification Trees." *Annals of Applied Statistics*, **3**, 1710–1737.

Malchow H (2008). *Political Targeting*. 2nd edition. Predicted Lists, LLC.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall, London.

Morgan JN, Sonquist JA (1968). "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association*, **58**, 415–434.

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

R Development Core Team (2011). "R: A Language and Environment for Statistical Computing." http://www.R-project.org/.

Rao CR, Toutenburg H (1997). *Linear Models: Least Squares and Alternative Methods*. 2nd edition. Springer-Verlag, New York.

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

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Wang Y, Witten I (1997). "Induction of Model Trees for Predicting Continuous Classes." In *Proceedings of the posters of the European Conference on Machine Learning*. University of Economics, Faculty of Informatics and Statistics, Prague, Czech Republic.

White H (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, **29**, 1–25.

Zeileis A (2005). "A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals." *Econometric Reviews*, **24**(4), 445–466.

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**(4), 488–508.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514.

Zhang H, Singer B (2010). *Recursive Partitioning and Applications*. 2nd edition. Springer-Verlag, New York.

Affiliation:

Thomas Rusch
Institute for Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien, Austria
E-mail: Thomas.Rusch@wu.ac.at

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://eeecon.uibk.ac.at/~zeileis/