Psychological Test and Assessment Modeling, Volume 61, 2019 (4), 389-417

Regression trees and random forests as alternatives to classical regression modeling: Investigating the risk factors for corporal punishment

Markus Fritsch[1], *Harry Haupt*[2], *Friedrich Lösel*[3] & *Mark Stemmler*[4]

Abstract

Examining the behavior of individuals is a challenging task due to complex patterns underlying the observable outcome. Even if all predictors contributing to the behavior are known in a particular context, the functional form and potential multi-level interactions between the predictors are difficult to grasp. Any modeling decisions involved in specifying a functional form should withstand comparisons with data-driven techniques. Decision trees and random forests enable data-driven modeling and are valuable tools to overcome limitations of least square regressions and validate existing results. We illustrate the relevant modeling steps required to carry out the two techniques by investigating the complex patterns of aggressiveness, dysfunctional parent-child interactions, and other risk factors for corporal punishment of children by their fathers. We replicate existing results on the corresponding risk factors, interpret the modeling outcomes, and describe the setting of relevant meta parameters in empirical practice.

Keywords: regression trees, random forests, functional form selection, predictor selection, corporal punishment, parenting behavior

[1] *Correspondence concerning this article should be addressed to:* Markus Fritsch | Department of Statistics, University of Passau, 94030 Passau, Germany; email: markus.fritsch@uni-passau.de
[2] Department of Statistics, University of Passau, 94030 Passau, Germany
[3] Institute of Criminology, Cambridge University, UK and University of Erlangen-Nuremberg, Germany
[4] Department of Psychology, University of Erlangen-Nuremberg, Germany

1 Introduction

Studies of risk factors for undesirable behavioral outcomes are key topics in psychology, medicine, criminology, and many other disciplines. Numerous methods aim to provide sufficiently valid results in practice. Depending on the respective research questions, the underlying research contains a broad range of methods, for example single items, structured clinical indices, or variations of the ordinary least square (OLS) model, i.e., multiple linear, binomial logistic, hierarchical, or multilevel approaches. OLS models clearly improve predictions because only the aggregation of multiple risk factors reach satisfactory validity (e.g., Bender & Lösel, 2005; Loeber & Farrington, 1998; Wallner, Lösel, Stemmler, & Corrado, 2018). However, the underlying OLS model may suffer from well-known problems, such as different scale levels of the included variables, non-normal distributions, heteroscedasticity across subgroups, influence of extreme scores, interaction effects, reproducibility of the results for different samples, and missing data. For these reasons, Haupt, Lösel, and Stemmler (2014) used Quantile Regression Analysis (QRA) as an alternative to OLS regression in a previous study and applied it to the issue of risk factors for corporal punishment in the parenting behavior of fathers. The findings illustrated the benefits of employing QRA. In particular, it turned out that various variables had different quantitative relations to the outcome at different score levels. Although the effect sizes were moderate, the study showed that it is helpful to apply more differentiated approaches to assess relations between risk factors and behavioral outcomes in psychology and other social sciences.

Basing on the previous research of Haupt et al. (2014), the present study describes other models that can reduce some of the problems of OLS techniques, in particular dealing with different scale levels and missing data. Furthermore, the methods address sequential/hierarchical decisions that are useful in practice. These are data-driven classification and regression trees (CART; see Breiman, Friedman, Olshen, & Stone, 1984). Classification trees are useful when the outcome is categorical, whereas regression trees should be used for continuous outcomes. As in other statistical models, one cannot avoid some problems. CART techniques may yield unstable results (meaning that the fitted model may exhibit pronounced changes when a small fraction of the input data is varied), low predictive performance, and do not allow standard statistical inference (for more extensive discussions of the properties of CART see Chapter 9.2.4 in Hastie, Tibshirani, and Friedman, 2009 or Chapter 8.1.4 in James, Witten, Hastie, and Tibshirani, 2013). Some of these disadvantages can, however, be overcome by averaging multiple trees. The model fit by this approach involves multiple trees called an ensemble of trees: Obtain the fitted value for the observation based on each of the individual trees and compute the arithmetic mean across all fitted values. Approaches which are based on variations of the idea of averaging multiple classification or regression trees are bagging (Breiman, 1996), boosting (Freund & Schapire, 1996, 1997; Friedman, 2002; Friedman, Hastie, & Tibshirani, 2000) and random forests (Breiman, 2001a). All of these techniques fit a model in data-driven fashion and can be a useful extension of the tools typically applied in psychological and social science research where replication and cross-validation is an

important problem (Lösel, 2018; Open Science Collaboration, 2015).

Against this background, the present article employs regression trees and random forests as statistical modeling techniques. Decision trees are practically highly relevant in sciences like psychology or criminology because diagnostic assessments often need a sequential strategy, for example in risk assessment of violent behavior (Monahan, 2012). Insofar, it is important to use the most relevant and well-replicated variables on a specific topic. Not rarely, the theoretical hypotheses on a specific topic are not consistent and the empirical impact of specific variables varies depending on the theoretical and empirical context of the respective investigation. Therefore, pursuing an exclusively "theory-driven" approach has limitations in practice. For example, many theories on criminal behavior only have moderate predictive validity, and have not been tested in comparison so that none really "failed" (Bernard, 1990). Furthermore, there is a deficit of integrative theories (Bruinsma, 2016; Lösel, 2017). For these and other reasons, practical decisions are often data- instead of theory-driven. From a scientific point of view this is not optimal, but daily practice requires decisions that may be more or less eclectic.

This article presents the above-mentioned methodological approaches that can reduce some problems of OLS studies. Our focus is on the description of the statistical models, but we also illustrate their content-oriented application by using outcome data of corporal punishment by fathers as in Haupt et al. (2014). In accordance with the question of replication, we expand the set of risk factors to assess the consistency of our findings.

In the first part of this article, we briefly describe the assumptions and algorithms of the above-mentioned two statistical models. Then we present the data for the empirical test and the respective results. Finally, we discuss the findings from a methodological perspective and draw conclusions about our content topic of parental corporal punishment.

2 Modeling framework, regression trees, and random forest models

The classical approach is to consider a general regression model

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a vector of dependent variables, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P)$ is a matrix that contains P predictors, and $\boldsymbol{\varepsilon}$ is a vector of idiosyncratic remainder components for all individuals $i = 1, \ldots, n$. The dependent variable is modeled as a function $f(\cdot)$ of the predictors.

Most commonly, a linear in parameters version of Equation (1) is chosen,

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \cdots + \boldsymbol{x}_P\beta_P + \boldsymbol{\varepsilon}, \tag{2}$$

where $\beta_1, \ldots, \beta_P$ denote the model parameters. For example, Haupt et al. (2014) assumed that a specific subset of the predictors $\boldsymbol{x}_1, \ldots \boldsymbol{x}_P$ and the linear functional form of the regression model in Equation (2) are given.

This stands in sharp contrast to a situation frequently encountered in empirical practice: Different theoretical hypotheses may exist, which cannot be nested in a regression

framework relying on a specific functional form, governing the interplay of a specific subset of predictors. In such a situation, regression trees and random forests avoid such specific choices and provide a data-driven alternative to model Equation (1). The next two subsections briefly explain the fundamentals of regression trees and random forests and describe the basis algorithms used for estimation. More details are provided by De'ath and Fabricius (2000) and Prasad, Iverson, and Liaw (2006) with a focus on ecological data, Miller, Lubke, McArtor, and Bergeman (2016) (for psychological data), Varian (2014) (for economic data), or in the vignette of the R-package rpart (Therneau & Atkinson, 2019). The third subsection discusses why regression trees and random forests provide a fruitful approach to analyze the risk factors for corporal punishment.

2.1 Regression trees

Decision trees can be fit with the CART (classification and regression tree) algorithm of Breiman et al. (1984). When the outcome is categorical, the decision trees are also referred to as classification trees; for continuous outcomes, *regression trees* is the customary term. Due to the (quasi-)continuous nature of the dependent variable considered in the empirical application, we describe the essential steps for fitting regression trees from a practical perspective.

Algorithm 1: Algorithm for fully growing a regression tree

Data: $(\boldsymbol{y}, \boldsymbol{X})$, with $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P)$
Result: Regression tree of depth d that partitions data into $R_{d,j}$ regions
Initialization: Set $s \leftarrow 0$, $d \leftarrow 0$, $q \leftarrow 0$, $q_d \leftarrow 0$, $j \leftarrow 1$, $J \leftarrow 1$, and $M_{R_{d,j}} \leftarrow 0$;
for *j=1,...,J* **do**
  set $q \leftarrow q + 1$;
  set $q_d \leftarrow q$;
  consider the fraction of the data $(\boldsymbol{y}_{d,j}, \boldsymbol{X}_{d,j})$ that fall into region $R_{d,j}$;
  for $p = 1, \ldots, P$ **do**
    at node q, assess the predictor $\boldsymbol{x}_{p,d,j}$, for a split $s_{d,j}$ according to the metric $\sum_{j=1}^{J} \sum_{i \in R_{d,j}} M_{R_{d,j}}$, with $M_{R_{d,j}} = M_{R_{d,j},\text{left}} + M_{R_{d,j},\text{right}}$ and $M_{R_{d,j},\text{left}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} < s_{d,j}\}, \quad M_{R_{d,j},\text{right}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} \geq s_{d,j}\}$;
  end
  choose the split $s_{d,j}$ that minimizes the metric $M_{R_{d,j}}$;
  compute $n_{d,j} = \sum_{i=1}^{n} I_{i \in R_{d,j}}$, for all $j = 1, \ldots, J$;
  if $M_{R_{d,j}} > M_{R_{d,j,min}}$ *and* $n_{d,j} > n_{d,j,min}$ **then**
    $q \leftarrow q + 1$;
  end
end
if $q > q_d$ **then**
  set $d \leftarrow d + 1$;
  set $J \leftarrow J + 1$ and continue looping over j (outer **for** loop above);
else
  return *the fully grown regression tree* T_J *of depth* d *with* J *terminal nodes*
end

Algorithm 1 illustrates the key steps required to fit a regression tree to the data $(\boldsymbol{y}, \boldsymbol{X})$ using the CART algorithm.

Note that s denotes a split, d the depth of the tree, q is a count for the number of nodes, q_d is a count for the number of nodes up to tree depth d, j indicates the region in which the data are separated, where the total number of regions is J, and $M_{R_{d,j}}$ is a metric calculated based on the observations in region R_j at tree depth d. In the first step of the algorithm, a regression tree of depth d is fully grown by recursive binary partitioning (or splitting) and the data are split into J non-overlapping regions $R_{d,1}, \ldots, R_{d,J}$ for a certain depth of the tree.

Algorithm 1 proceeds in a top-down and greedy fashion. The former means that at the first node at the top of the tree, which is also referred to as root node, all observations are considered. The latter refers to the way the data are split: At each of the q nodes (i.e., in a particular region $R_{d,j}$), only the reduction in a pre-specified measure $M_{R_{d,j}}$ is considered and all previous and later steps (or splits) are not incorporated into the decision about the split. In formal terms, splitting the space spanned by the predictors $\boldsymbol{X}$ at tree depth d involves partitioning the predictors $\boldsymbol{X}$ into J high-dimensional rectangles in order to minimize the metric

$$M_{R_{d,j}} = M_{R_{d,j},\text{left}} + M_{R_{d,j},\text{right}}, \tag{3}$$

with $M_{R_{d,j},\text{left}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} < s_{d,j}\}$ and $M_{R_{d,j},\text{right}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} \geq s_{d,j}\}$. In Equation (3), $M_{R_{d,j}}$ in node q (where a split in region $R_{d,j}$ is considered) represents a metric that is composed of the sum of an analogous measure on the left of the split $M_{R_{d,j},\text{left}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} < s_{d,j}\}$ and on the right of the split $M_{R_{d,j},\text{right}} = \{\boldsymbol{X} | \boldsymbol{X}_{d,j} \geq s_{d,j}\}$. When fitting regression trees, a popular choice of $M_{R_{d,j}}$ is the sum of squared residuals in region $R_{d,j}$ ($SSR_{R_{d,j}}$) (see, e.g., James et al., 2013), such that

$$SSR_{R_{d,j}} = \sum_{i \in R_{d,j}} (y_i - \hat{y}_{R_{d,j}})^2, \tag{4}$$

where $SSR_{R_{d,j},\text{left}} = \sum_{i \in R_{d,j},\text{left}} (y_i - \hat{y}_{R_{d,j},\text{left}})^2$ and $SSR_{R_{d,j},\text{right}} = \sum_{i \in R_{d,j},\text{right}} (y_i - \hat{y}_{R_{d,j},\text{right}})^2$. Note that splitting continues, until the improvement in the metric indicated in Equation (4) is below a pre-specified minimum value $M_{R_{d,j},\text{min}}$. Alternatively, the splitting process is stopped, when the number of observations in a particular node of the tree does not exceed a pre-specified minimum $n_{d,j,\text{min}}$ (see, e.g., Hastie et al., 2009). Per an adjustable default, the `R` implementation of the CART algorithm `rpart`, for example, only considers splits that improve the multiple $r^2_{d,j}$ (derived from the $SSR_{R_{d,j}}$) in region $R_{d,j}$ by a minimum of $M_{R_{d,j},\text{min}} = 0.01$, when at least $n_{d,j,\text{min}} = 20$ observations are available in a particular region $R_{d,j}$ (see Therneau & Atkinson, 2019). The nodes where the data cannot be split anymore are called terminal nodes (or leafs) and the result from the algorithm is a fully grown regression tree T_J with J terminal nodes.

The second step of the CART algorithm involves fitting a constant to each dependent

variable within region $R_{d,j}$. The fitted value is given by the arithmetic mean

$$\hat{c}_j = \frac{1}{\sum_{i=1}^{n} I(x_i \in R_{d,j})} \cdot \sum_{i \in R_{d,j}} y_i, \tag{5}$$

where the indicator function counts all observations that fall into region $R_{d,j}$ (see, e.g., Hastie et al., 2009). The resulting fully grown regression tree T_J can be related to the bias-variance trade-off as follows: A large tree possesses low bias, as it fits the training data reasonably well. However, when a certain fraction of the sample changes, a large tree frequently exhibits instability of its structure and the resulting predictions (see, e.g., James et al., 2013; Kuhn & Johnson, 2013). To account for this aspect and reduce the risk of overfitting, some of the internal nodes (i.e., non-terminal nodes) of T_J are typically removed after fully growing the tree. This is referred to as pruning and reduces the number of terminal nodes and, hence, decreases the model complexity. Breiman et al. (1984) suggest a pruning procedure, which involves a cost complexity criterion that balances the trade-off between goodness of fit to the data and model complexity. Slightly modifying the notation in Hastie et al. (2009), such a criterion may be defined as

$$E_\alpha(T_L) = \sum_{j=1}^{L} SSR_j + \alpha L, \tag{6}$$

where $T_L \subseteq T_J$ denotes a regression tree with $L \leq J$ terminal nodes nested in the fully grown regression tree T_J. The parameter α in Equation (6) denotes a tuning parameter that governs the trade-off between the goodness of fit to the training data and the model complexity measured by the number of terminal nodes of the tree.

The two components of Equation (6) are: The residual sum of squares SSR_j of a regression tree with j terminal nodes. This term measures the goodness of fit and decreases as the fit to the data increases. The second component is a penalty term that depends on the parameter α and the number of terminal nodes L. Note that the penalty term increases as L increases. Fully grown regression trees are pruned such that Equation (6) is minimized.

The tree size depends on α as follows: For a large value of α (i.e., $\alpha \to \infty$), the tree consists of only the root node, while a small value (i.e., $\alpha \to 0$) corresponds to the fully grown regression tree. Therneau and Atkinson (2019) rescale α and label the rescaled parameter cp, where $cp \in [1; M_{R_{d,j},\text{min}}]$: Setting cp to the former value results in the tree that contains only the root node and the latter results in the fully grown regression tree.

Each tree size corresponds to a specific value of α in Equation (6). In practice, pruning is conducted such that for every internal node of the tree, the effect of collapsing the respective node (i.e., reducing the tree size) is assessed by a suitable criterion. A typical choice is to employ a criterion based on out-of-sample errors, which can be calculated using K-fold cross-validation. This approach randomly partitions the data into K

subsamples of equal size. A model is then fit using $K-1$ subsamples as training data and the remaining subsample to calculate the out-of-sample errors and the chosen criterion. Then, repeat and average over each of the K folds. The decision rule is to delete the internal node of the tree which yields the minimum increase in the chosen out-of-sample error criterion. This is done until all internal nodes of the tree are deleted and the tree consists of the root node only. Cost complexity pruning results in a sequence of (nested) subtrees and a corresponding out-of-sample error criterion for each of the trees, which can be characterized by the two components in Equation (6) (see, e.g., Hastie et al., 2009). The final step of the pruning process is to identify the sub-tree T_L that exhibits the lowest value for the criterion (or the sub-tree within one standard error of this tree; see, e.g., Breiman et al., 1984, James et al., 2013, or Kuhn and Johnson, 2013).

2.2 Random forests

Bagging (short for bootstrap aggregation; Breiman, 1996) regression trees refers to drawing $b = 1, \ldots, B$ bootstrap samples of identical size n_{boot} from the original training data $(\boldsymbol{y}, \boldsymbol{X})$ with replacement and fully growing one regression tree T_{J_b} to each of the B individual bootstrap samples according to Algorithm 1. Then, equal weights are assigned to all B individual regression trees and the model fit by the bagging procedure is a collection (or ensemble) of regression trees. The prediction made by the ensemble is the arithmetic mean of the predictions obtained from the B trees fit to each individual bootstrap sample.

The random forest approach of Breiman (2001a) is basically a subtle extension of bagging used to stabilize the properties of regression trees. Randomness is introduced when constructing the individual trees for the $b = 1, \ldots, B$ bootstrap samples: Instead of considering all P predictors at each node of each tree (in order to assess if the data should be split into two further subgroups by recursive binary partitioning), the algorithm only considers a subset m_{try} of the predictors, with $m_{\text{try}} < P$. Similar to bagging, the prediction made by a random forest – which is basically an ensemble of regression trees – is the simple average of the predictions obtained from the individual trees (Breiman, 2001a). A random forest algorithm is displayed in Algorithm 2.

Algorithm 2: Algorithm for fitting a random forest

Data: $(\boldsymbol{y}, \boldsymbol{X})$, with $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P)$
Result: Ensemble of regression trees
Initialization: Set the number of bootstrap samples B;
for $b = 1, \ldots, B$ **do**
  draw a bootstrap sample (with replacement) of size n_{boot} from $(\boldsymbol{y}, \boldsymbol{X})$;
  fit a fully grown regression tree T_J to $(\boldsymbol{y}, \boldsymbol{X})$ according to Algorithm 2, while considering only $m_{\text{try}} < P$ of the predictors in each region $R_{d,j}$;
end
return *the (bootstrap) ensemble of trees* $\{T_{J_b}\}_1^B$

The model fit by the algorithm is

$$\hat{f}_{\text{rf}}^{B}(\boldsymbol{X}) = \frac{1}{B} \cdot \sum_{b=1}^{B} T_{J_b}(\boldsymbol{X}). \tag{7}$$

With respect to the predictive performance, we have to consider the following. Since individual trees provide predictions that have low bias and high variance, introducing randomness when fitting the individual regression trees often improves the predictive performance substantially. This improvement results from a combination of two different techniques that reduce the variance of the prediction made by the ensemble in Equation (7), while introducing a slightly higher bias compared to the situation when fitting a single regression tree: First, averaging over the individual regression trees (fit on bootstrap samples drawn from the original data with replacement) reduces the variance. Second, only a subset of the predictors at each node of each tree is considered when assessing if the data should be split into two (additional) subgroups. This further reduces the variance by de-correlating the B bootstrap regression trees fit by the algorithm (see, e.g., Hastie et al., 2009).

2.3 Application of the models to data on corporal punishment

As in the previous study of Haupt et al. (2014), we chose fathers' corporal punishment in parenting for an empirical application of the statistical models. This is insofar an adequate example, as there are controversial views on this topic in the scientific literature. Negative effects of serious, repeated corporal punishment and physical abuse of children are well documented in the literature (for an overview see Bender & Lösel, 2015). There are different forms of legal bans (e.g., penal or civil law) of corporal punishment in various countries like Sweden, Austria and Germany aimed to reduce escalations to these serious forms and some authors recommend a general ban of spanking (Afifi et al., 2017). Perhaps due to these and other arguments physical punishment decreased substantially in Western countries. However, there is less agreement about the detrimental effects of mild forms of corporal punishment like occasional slapping. One branch of the literature advocates that corporal punishment is generally detrimental to child development as it enhances the risk of long-term outcomes like antisocial behavior, cognitive, and emotional distress (e.g., Gershoff, 2002, 2010; Straus, 2009, 2010). However, in a meta-analysis of longitudinal studies on the effects of spanking and corporal punishment on children's externalizing, internalizing, or cognitive problems the effect sizes were very small or partially not significant (Ferguson, 2013). Various authors emphasize that mild and occasional forms of corporal punishment in an intact parent-child relationship are not harmful to child development and can be effective measures of discipline – where the primary goal is to set boundaries and not to exercise power (Baumrind, Larzelere, & Cowan, 2002; Larzelere & Baumrind, 2010; Scarr & Deater-Deckard, 1997). Conflicting hypotheses not only exist about the consequences, but also about the risk factors for corporal punishment in parenting. The potential risk factors include evolutionary and cultural framing conditions, demographic family variables, strain and stressors in the family, parental personality

factors, mental health problems, characteristics of the children, a lack of protective factors in the social network and other features (for an overview see Bender & Lösel, 2015). Some of these risk factors are, for example, the socio-economic status of the family, intergenerational transmission of physical punishment in families, generally punitive methods of discipline in the family, family- and work-related stress, and a lack of external help in dealing with parenting problems (e.g., Ellonen, Peltonen, Pösö, & Janson, 2017; Masuda, Lanier, & Hashimoto, 2019; Peltonen, Ellonen, Pösö, & Lucas, 2014; Seay, Jahromi, Umaña-Taylor, & Updegraff, 2016; Widom, Czaja, & DuMont, 2015). However, most of these and other risks are discussed controversially in the literature, e.g., with regard to the strengths of relations, research designs, seriousness of the predictor or outcome variables, and measurement issues. As mentioned above, divergent assumptions are a key issue for data-driven decision trees and random forests. Therefore, we use our data on this topic to determine the relevant predictors, their degree of interaction, and the functional form by applying these models in a data-driven approach.

3 Method

3.1 Sample and data description

The data set employed to illustrate regression trees and random forests is a subsample taken from a representative sample collected in the Erlangen-Nuremberg Development and Prevention Study (Lösel, Stemmler, & Bender, 2013; Lösel, Stemmler, Jaursch, & Beelmann, 2009). We used data employed by Haupt et al. (2014) extended by additionally available predictors to study the risk factors for corporal punishment of elementary school children by their fathers. The data set contained 675 observations, where 199 of the observations had one or more missing values. We omitted all entries with missing values in any of the predictors or the response variable and only used the $n = 476$ complete lines in the data set for our analysis[1].

Details on the data such as the percentage of biological fathers and mothers, marital status of the parents, and national identity of the child are provided in Haupt et al. (2014). The computations in this paper were carried out in `R` version 3.5.3 (R Core Team, 2019). We used functions implemented in the packages `rpart` (Therneau & Atkinson, 2019) and `party` (Hothorn, Hornik, Strobl, & Zeileis, 2019) to fit and visualize regression trees, and functions from the package `randomForest` (Liaw & Wiener, 2018) to estimate random forests.

Corporal punishment was our dependent variable. As Haupt et al. (2014), we used the fathers' self-report in a German version of the Alabama Parenting Questionnaire (APQ; Shelton, Frick, & Wootton, 1996). The scores could range from 1 to 3, where large values indicate a high level of physical punishment. While Haupt et al. (2014) only included eight risk factors for corporal punishment in their analysis, we considered $P = 26$

[1] Note that an alternative is to impute the missing values by using the observed data, secondary data, or regression-based approaches. For various approaches of multiple imputation see Kleinke (2018) and Kleinke, Stemmler, Reinecke, and Lösel (2011).

potential predictors. As mentioned above, the influence of corporal punishment may vary according to other family characteristics. Therefore, we included demographic data, family interactions, other parenting variables, and fathers' personality. In the following, we briefly describe the various potential predictors:

Other parenting behavior (beyond corporal punishment) was captured by items from the APQ (Shelton et al., 1996). The measures used in the Erlangen-Nuremberg Development and Prevention Study were *inconsistent discipline*, *other disciplinary practices*, *monitoring and supervision*, *parental involvement*, and *positive parenting*. All measures were calculated by aggregating questions answered by the parent on a five point scale ranging from 'never' (lowest) to 'always' (highest).

Dysfunctional parent-child interaction was measured by the Parenting Stress Index (PSI; Abidin, 1995) and its scales on *parental distress*, *dysfunctional parent-child interaction*, and *difficulty of child*. The scores were calculated by aggregating the scores on a five point scale ranging from 'strongly agree' (lowest) to 'strongly disagree' (highest).

Father's personality was assessed by the revised version of the Freiburg Personality Inventory questionnaire (FPI-R; Fahrenberg, Hampel, & Selg, 1989). From this questionnaire we used the fathers' scores *aggressiveness*, *life satisfaction*, *social orientation*, *achievement orientation*, *inhibition*, *excitability*, *stress*, *physical complaints*, *health concerns*, *candor*, *extraversion*, and *emotionality*.

Demographics were captured by the measures *socioeconomic status* (Geißler, 1994), *age of the father*, the *age and age difference of the parents*, and the *gender* of the elementary school child.

Table 1 contains a brief description of the variables employed in our analysis. Further details on the data set are available in Haupt et al. (2014).

Table 1:
Variable description for the dependent variable $APCP$ and the $P = 26$ predictor variables contained in the Erlangen-Nuremberg Development and Prevention Study data ($n = 476$).

variable name	description
OLDER.M	indicator if mother is older than father
OLDER.F	indicator if father is older than mother
FEMALE	indicator if child is female
AGE	age of father
AGE.DIFF	absolute age difference of parents
SES	measure for socioeconomic status
APIN	measure for parental involvement
APPP	measure for positive parenting
APMS	measure for monitoring and supervision
APID	measure for inconsistent discipline
APCP	measure for corporal punishment (dependent variable)
APOD	measure for other dicipline practices
PIPD	measure for parental distress
PIDI	measure for dysfunctional parent-child interaction
PIDC	measure for difficulty of child
FPAG	measure for aggressiveness of father
FPLS	measure for life satisfaction
FPSO	measure for social orientation
FPLO	measure for performance orientation
FPGH	measure for inhibition
FPER	measure for excitability
FPBS	measure for stress
FPKB	measure for physical complaints
FPGS	measure for health concerns
FPOF	measure for candor
FPEX	measure for extroversion/conviviality
FPEM	measure for emotionality

3.2 Descriptives and exploratory data analysis

Of the $n = 476$ elementary school children included in the data, 89.3 percent were the only child of the family participating in the study, while 10.7 percent of the families had more than one child. About 50.1 percent of the children were female and the other half were male. The age of the father ranged from 27 to 59 years (mean age $\bar{x}_{\text{age},f} = 39$ years), while mothers' age was between 23 and 49 years ($\bar{x}_{\text{age},m} = 37$ years). In 74.6 percent of the families the father was older than the mother, in 16.1 percent the mother was older, and in 9.3 percent both parents were of the same age.

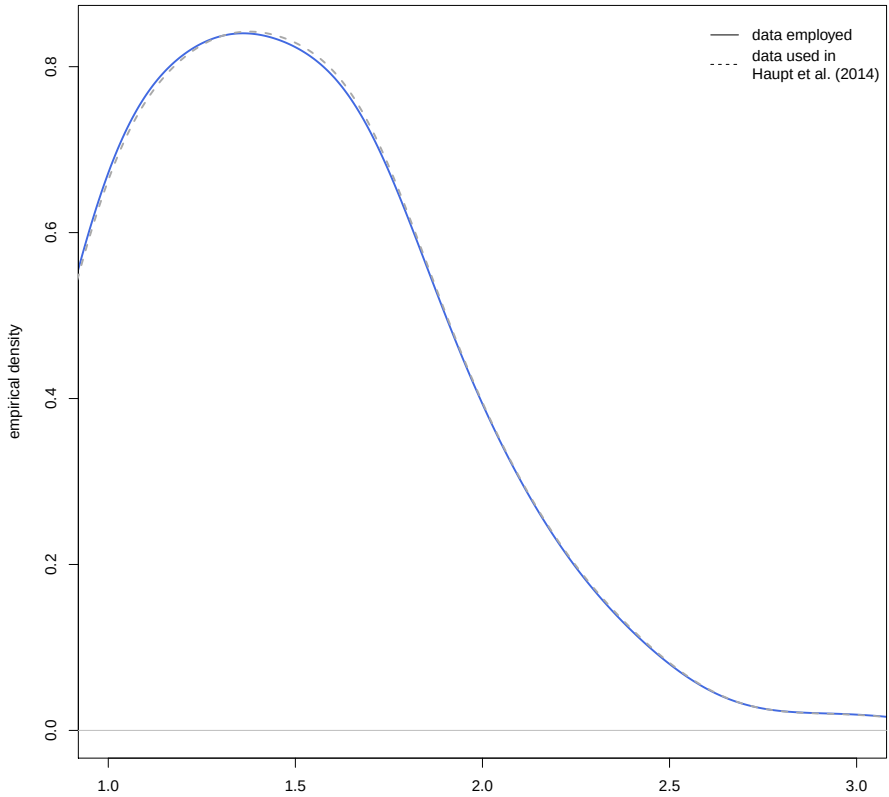
Table 2 shows the five number summary of Tukey extended by the mean for the 23 continuous predictor variables included in the data set and the measure for *corporal punishment* ($APCP$) used as dependent variable.

Table 2:
Five number summary of Tukey extended by the mean for the dependent variable $APCP$ and the 23 continuous predictor variables contained in the Erlangen-Nuremberg Development and Prevention Study data ($n = 476$).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AGE	27.00	36.00	38.00	39.20	42.00	59.00
AGE.DIFF	0.00	1.00	2.00	3.14	4.50	23.00
SES	1.30	2.05	2.30	2.30	2.65	3.00
APIN	1.80	2.90	3.30	3.33	3.70	4.70
APPP	2.00	3.50	3.83	3.85	4.17	5.00
APMS	1.00	1.50	1.90	1.93	2.30	3.40
APID	1.00	2.00	2.33	2.38	2.67	3.83
APCP	1.00	1.00	1.33	1.49	1.67	3.00
APOD	1.00	2.00	2.29	2.37	2.71	3.57
PIPD	12.00	18.00	22.00	22.83	27.00	46.00
PIDI	12.00	15.00	18.00	19.12	22.00	38.00
PIDC	12.00	19.00	23.00	23.48	27.00	44.00
FPAG	0.00	2.00	3.00	3.64	5.00	12.00
FPLS	0.00	7.00	9.00	8.52	10.00	12.00
FPSO	0.00	5.00	7.00	6.80	9.00	12.00
FPLO	0.00	6.00	8.00	7.63	10.00	12.00
FPGH	0.00	2.00	4.00	4.68	7.00	12.00
FPER	0.00	3.00	4.00	4.72	6.00	12.00
FPBS	0.00	4.00	6.00	6.04	9.00	12.00
FPKB	0.00	0.00	1.00	1.66	2.00	9.00
FPGS	0.00	2.00	4.00	4.08	6.00	11.00
FPOF	0.00	4.00	6.00	6.08	8.00	12.00
FPEX	0.00	4.00	7.00	6.66	9.00	14.00
FPEM	0.00	2.00	4.00	4.40	7.00	13.00

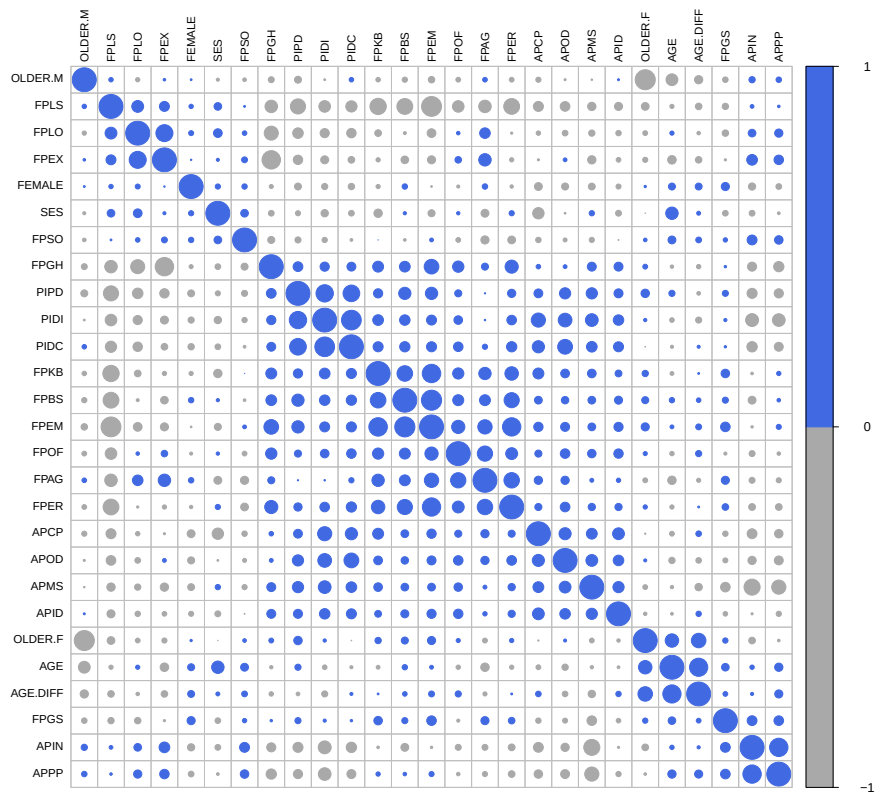
Figure 1 shows the similarity of the empirical density of the dependent variable $APCP$ as used in this paper (solid blue line), and the empirical density of $APCP$ as employed in Haupt et al. (2014) (dashed grey line). Note that high values of the dependent variable indicate more severe corporal punishment.

Figure 1:
Empirical density of *corporal punishment* ($APCP$) of elementary school children by their fathers as used in the empirical application (blue) and the corresponding density of the data set employed in Haupt, Lösel, and Stemmler (2014). The sample size n is 476 for the former version of the data set and 485 for the latter. The abscissa of the plot reflects the range of the dependent variable.



Similar to Haupt et al. (2014), we investigated the pairwise correlations of the dependent variable $APCP$ and all $P = 26$ predictor variables contained in the data set. Figure 2 visualizes these pairwise correlations, where positive correlations are indicated as blue circles, negative correlations as grey circles, and the strength of the absolute correlation is indicated by the size of the circle.

Figure 2:
Plot of the pairwise correlations of the dependent variable $APCP$ and the $P = 26$ predictor variables contained in the Erlangen-Nuremberg Development and Prevention Study data ($n = 476$). Blue circles indicate a positive pairwise correlation, grey circles represent a negative pairwise correlation, and the size of the circle shows the strength of the pairwise correlation (a large circle indicates a large absolute correlation and vice versa; the absolute correlations in the data set range from 0.0002 to 0.7).



4 Modeling results

As suggested by Miller et al. (2016), we standardized the dependent variable and the predictors, so that all variables in the data set possessed a mean of zero and a standard deviation of one. We applied the function `rpart` (from the same-named R-package) to a training data set of sample size $n_{\text{train}} = 381$ taken from the complete set of $n = 476$ observations in the Erlangen-Nuremberg Development and Prevention Study data

(roughly 80 percent of observations). Note that throughout the empirical application we held out the remaining 20 percent of observations as a test data set ($n_{\text{test}} = 95$) to compute out-of-sample error measures via the validation set approach after having fit the models[2]. This means that we used only the n_{train} observations to fit the models and used the n_{test} observations to assess their out-of-sample performance.

4.1 Regression trees

Figure 3 illustrates the resulting fully grown regression tree based on our training data. The splitting criterion is indicated at every node and the splitting rules are shown at the branches of the tree. At the root node, for example, the data are split according to whether the score for the *dysfunctional parent-child interaction* is below 0.46. The first internal node following the root node on the left-hand side splits the data according to whether the score for *other disciplinary measures* is below or above -1, where a score below -1 leads to a terminal node. For the $J = 20$ terminal nodes, Figure 3 also indicates the fractions of the data falling into the respective regions. The terminal node on the left-hand side of the figure, for example, contains 15 percent of the training data and the mean of the dependent variables falling into that region (and, hence, the fitted value) is -0.63.

To avoid overfitting the training data, we carried out cost-complexity pruning of the fully grown regression tree in Figure 3 by K-fold cross-validation and use $K = 10$ (the default of the corresponding function in the `rpart` package). Figure 4 illustrates a relative out-of-sample error criterion obtained from K-fold cross-validation (as described in Section 2.1) on the ordinate and the number of terminal nodes of the tree (Therneau & Atkinson, 2019) and on the upper and lower abscissa, respectively. The relative out-of-sample error criterion is the mean square error (mse) of the considered sub-tree T_L with L terminal nodes relative to the mse of the tree, where the only terminal node is the root node.

[2] We chose an 80/20-split of the data to have a sufficiently large training data set, as fitting random forests requires tuning a meta parameter. We selected the meta parameter based on 10-fold cross-validation based on the training data.

Figure 3:
Fully grown regression tree for the standardized Erlangen-Nuremberg Development and Prevention Study data ($n_{\text{train}} = 381$). All nodes in the plot contain information on the variable used to split the data. The top node of the tree is also referred to as root node, the nodes where the data cannot be split anymore are labelled terminal nodes (or leafs), and all other nodes are referred to as internal nodes. The terminal nodes also contain the fraction of the observations that fall into each of the regions of the predictor space and the corresponding mean of the dependent variable $APCP$. The splitting criteria are given at the lines connecting the nodes (or branches).

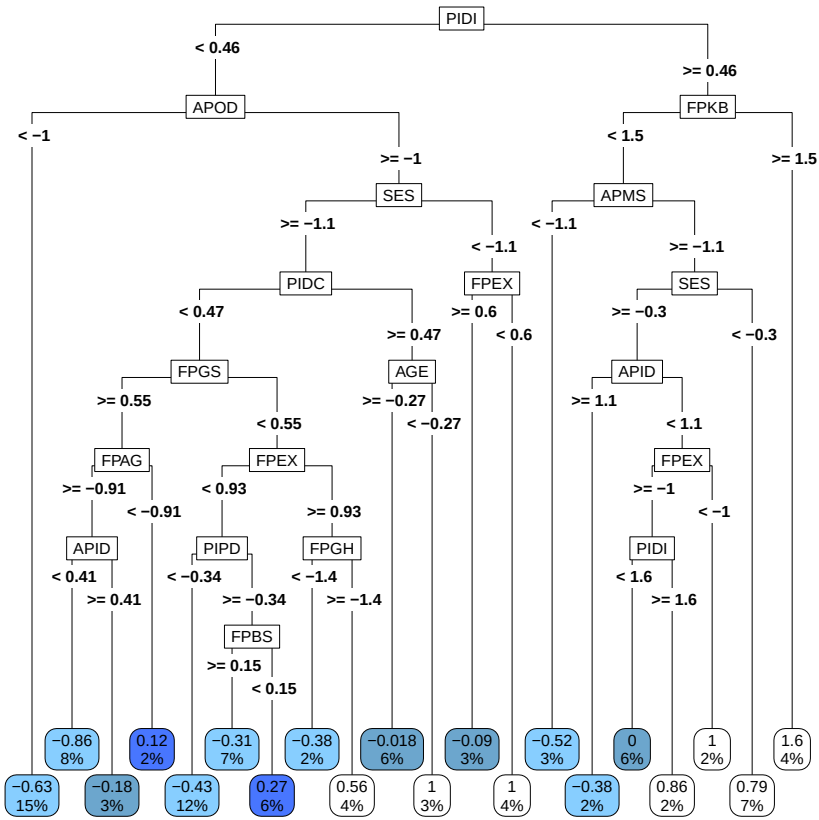
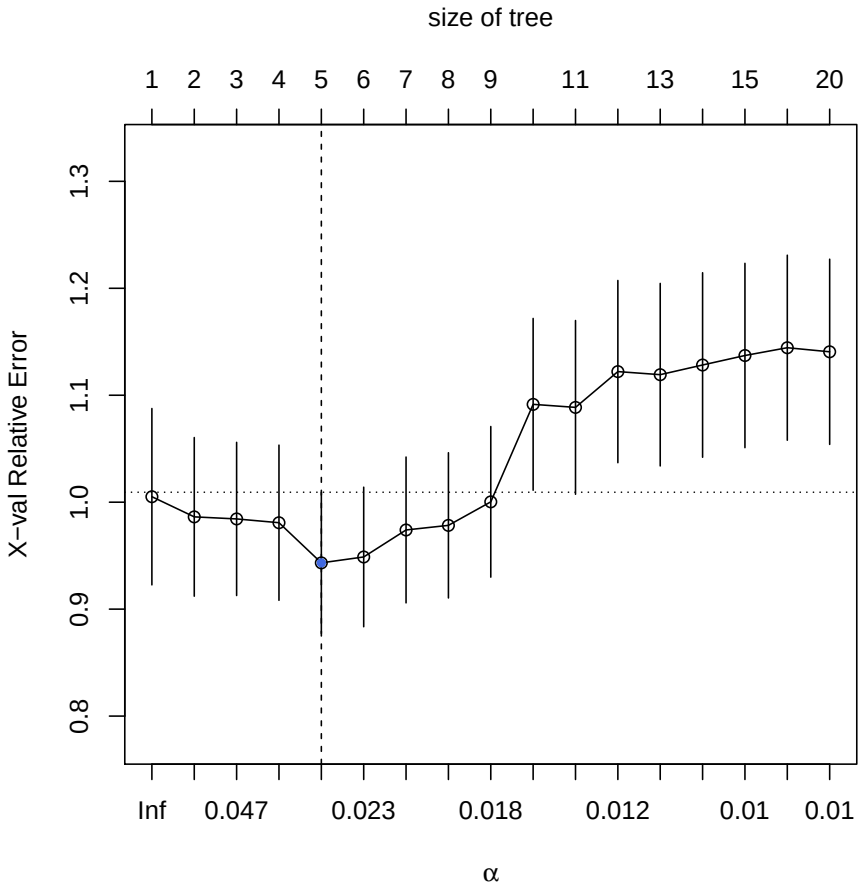


Figure 4 allows to derive conclusions where the regression tree may be pruned. In the case illustrated, the minimum relative mse obtained from 10-fold cross-validation results for a regression tree with five terminal nodes.

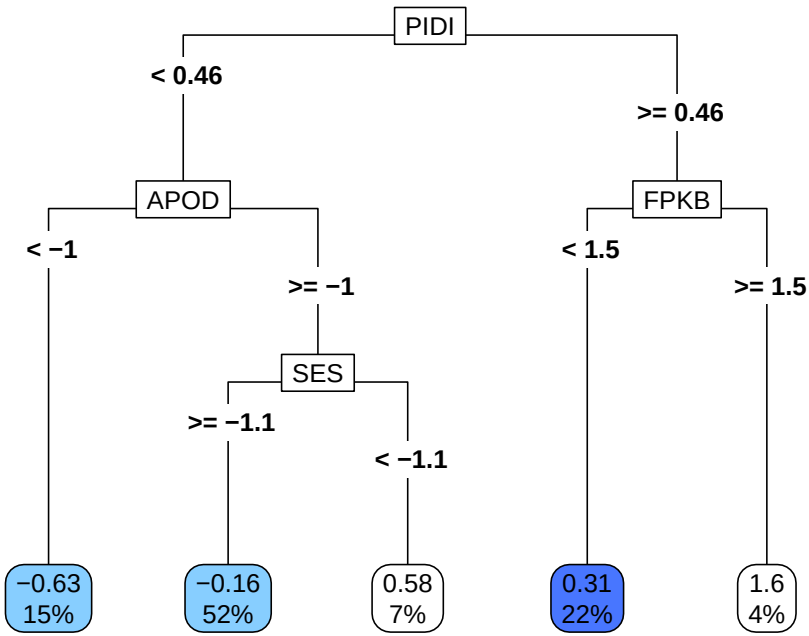
Figure 5 displays the pruned regression tree resulting from reducing the number of terminal nodes to five – as suggested by the cost-complexity diagnostics in Figure 4.

Figure 4:
Cross validated relative out-of-sample mse (ordinate) plotted against the tuning parameter α (lower abscissa) and the terminal nodes of the regression tree (upper abscissa). The tree with $J = 20$ terminal nodes corresponds to the fully grown regression tree, while the tree with $J = 1$ terminal nodes is a tree which contains only the root node. For each number of terminal nodes, the subtree with minimum 10-fold cross-validated mse obtained by cost complexity pruning is shown in the plot. The ordinate shows the out-of-sample mse of a tree with a particular number of terminal nodes relative to the tree which contains the root node only.



An alternative visualization of the cost-complexity pruned regression tree according to Hothorn et al. (2019) is shown in Figure 6. Compared to Figures 3 or 5, the visualization of the resulting tree in Figure 6 comprises additional information: The box plots for the five terminal nodes visualize the five number summary of Tukey for the dependent variable at the respective terminal node and the corresponding number of observations contained in each node. This effectively illustrates that the median score of *corporal*

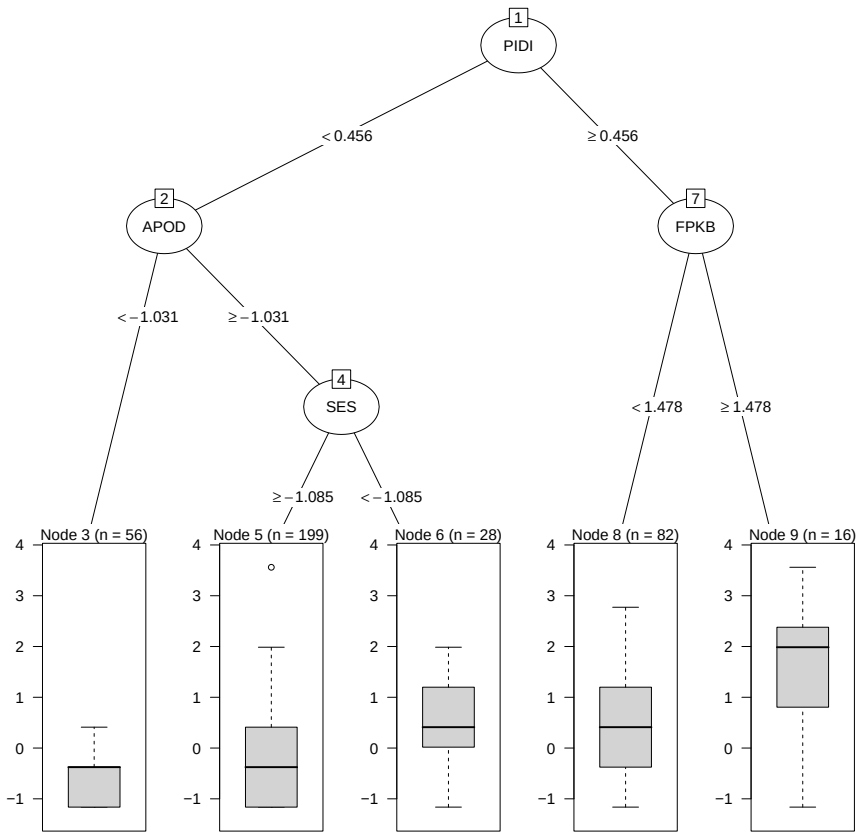
Figure 5:
Regression tree resulting when cost-complexity pruning the fully grown regression tree in Figure 3 according to the minimum relative mse from the 10-fold cross-validation shown in Figure 4.



punishment of elementary school children by their fathers increases from left to right and that the corresponding inter-quartile range (i.e., the size of the box) varies across the five terminal nodes.

Note that the cost complexity pruned regression tree shown in Figures 5 and 6 exhibits substantial differences to the multiple linear regression estimated in Haupt et al. (2014), as both models employ different predictors and none of the predictors overlap. However, both models yield a comparable in-sample fit (0.223 for the regression tree-based model compared to 0.251 for the multiple linear regression). For computing out-of-sample error criteria for the predictive performance of the models, we used the validation set approach and employed 80 percent of the data to train fully grown trees and prune

Figure 6:
Alternative visualization of the cost-complexity pruned fully grown regression tree from Figure 3. Besides the information on the split and the number of observations that fall into each terminal node, the plot shows boxplots for the dependent variable in each of the five resulting terminal nodes.

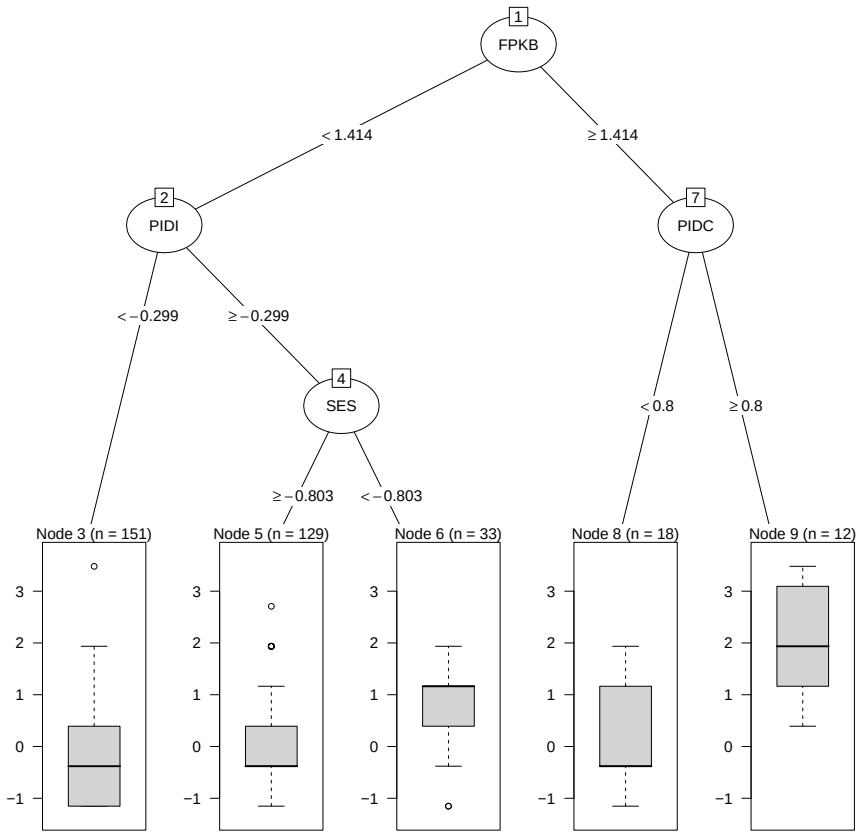


the trees back by $10-$fold cross validation (analogous to the trees fit in this section). We then use the remaining 20 percent of the data to obtain the $rmse$ for both models. The regression tree model has $rmse = 0.982$ ($mse = 0.964$, $mae = 0.762$) and the multiple linear regression model reported in Haupt et al. (2014) yields $rmse = 0.979$ ($mse = 0.959$, $mae = 0.771$). Though the structure of the two models is different – while the multiple linear regression model approximates the underlying relationship by a functional form that is linear in parameters, regression trees provide an approximation based on interactions of the partitioned predictor space – both models possessed a similar

in-sample fit and yielded a comparable out-of-sample performance. The multiple linear regression model seemed to be a valid approximation of the underlying process when compared to regression trees.

Regression trees are frequently criticized due to low predictive performance and lack of stability when the underlying training data set exhibits small changes (Kuhn & Johnson, 2013). In Figure 7, we exemplarily illustrate some instability in our data. For this purpose, we randomly deleted ten percent of observations, fully grew a regression tree based on the remaining data and subsequently carried out cost-complexity pruning.

Figure 7:
Cost-complexity pruned regression tree obtained after deleting 10% of observations from the training data at random.



The resulting regression tree contains a different root node and two further different

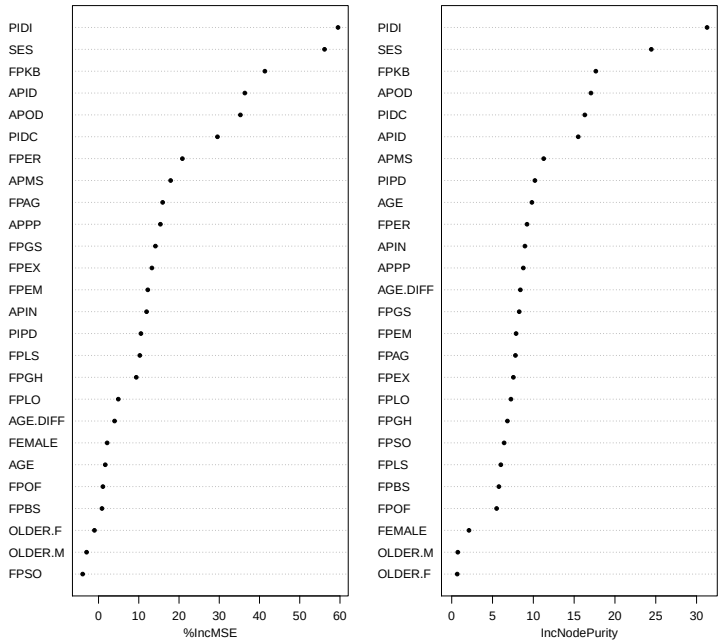
nodes compared to the tree shown in Figures 5 and 6. This illustrates the aforementioned instability of regression trees when small fractions of the input data used to grow a tree vary.

4.2 Random forests

The meta parameters for fitting random forests illustrated in this paper were fixed as follows: The number of bootstrap regression trees was set to $B = 5000$. No more splits in the individual trees were considered when the number of observations in a node was below $n_{d,j,\text{min}} = 20$ (which is identical to the value chosen to fit the individual trees). The number of predictors considered to split the data at each node was chosen by $10-$fold cross-validation and set to $m_{\text{try}} = 13$. The in-sample fit of this random forest is 0.208 and lies below both our regression tree model and the multiple linear regression model of Haupt et al. (2014). Concerning the out-of-sample error measures $rmse = 0.961$ ($mse = 0.924$, $mae = 0.773$), random forests yielded an improvement in the predictive performance with respect to $rmse$ and mse compared to the two other models.

Since random forests are ensembles of regression trees and due to the construction of the individual trees of a random forest (i.e., by considering subsamples of the training data and by not including all predictors when the data are assessed for splits), the predictors selected to split the data vary across the individual trees and the model structure of the full ensemble is difficult to grasp. One way to make sense of the model structure is to visualize the individual trees similarly to Figures 4 and 6. Eyeballing (possibly) many visualizations of tree structures might, however, not be the most appropriate way to gain an intuition about the model structure. Figure 8 shows a more suitable alternative. The figure indicates how often the predictors are chosen as splitting variables within the tree fitting process and is typically interpreted as a measure to assess the importance of a predictor (i.e., the more frequent a predictor is chosen to split the training data, the more important is the respective predictor). In the figure, the variables are sorted from the (statistically) most important variable (top) to the least important variable (bottom) according to a particular criterion. The left part of the figure assesses the importance of the variables with the percentage increase in mse when randomly permuting the respective predictor variable compared to the case when the predictor is not permuted. The right part of the figure shows the increase in SSR. The mse was computed on the basis of the out-of-bag observations (i.e., the observations that are left out at each bootstrap iteration), while the SSR is based on the observations used to fit the respective tree at each bootstrap iteration. According to Figure 8, *socioeconomic status* (SES), *physical complaints* ($FPKB$) *inconsistent discipline* ($APID$), *dysfunctional parent-child interaction* ($PIDI$), *difficulty of child* ($PIDC$) and *other discipline practices* ($APOD$) were the main risk factors for corporal punishment by fathers.

Figure 8:
Variable importance plot for the random forest fit to the Erlangen-Nuremberg Development and Prevention Study data. The number of randomly selected predictors at each node to determine potential split points $m_{\text{try}} = 13$ is chosen by 10-fold cross-validation. The importance of the variables is assessed according to the percentage increase in an out-of-bag criterion (mse; left-hand side) and the absolute increase in an in-sample criterion (SSR; right-hand side) when randomly permuting one particular predictor variable. The predictors are sorted from most important (top) to least important (bottom).



The modeling outcome differs when considering random forests and standard parametric methods like multiple linear regression. While the latter allows the researcher to derive marginal effects, standard inference, and conduct ceteris paribus interpretation, this is not the case for random forests. One way to assess the partial effect of a predictor on the dependent variable is shown in Figures 9 and 10. The figures illustrate the predictor on the abscissa and the effect of the predictor on the dependent variable on the ordinate. The dashes at the abscissa of the plots indicate the sample deciles of the respective predictor variable. The plots illustrate that the partial effects of the predictors may not be constant across the whole conditional distribution of the dependent variable. Consider, for example, the predictor SES shown in Figure 10: The partial effect was positive, but decreased with an increase in socioeconomic status for low to middle levels of the predictor; for middle to high levels of SES, the effect was negative and roughly constant. For some of the other predictors similar patterns (i.e., non-constant partial effects on the dependent variable) occurred. Overall, the partial effects implied by the random forest

of regression trees indicated that the marginal effects on the dependent variable may be nonlinear and depend on the level of the predictor variable.

Figure 9:
Partial effect of the predictor *dysfunctional parent-child interaction* ($PIDI$) (abscissa) on the dependent variable (ordinate) implied by a random forest of regression trees. The rugs on the abscissa indicate the deciles of the predictor variable.

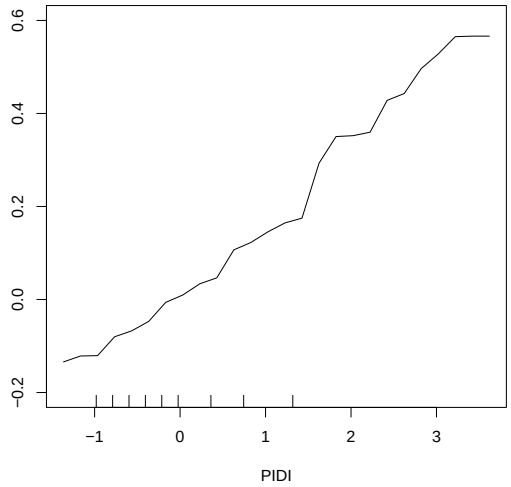
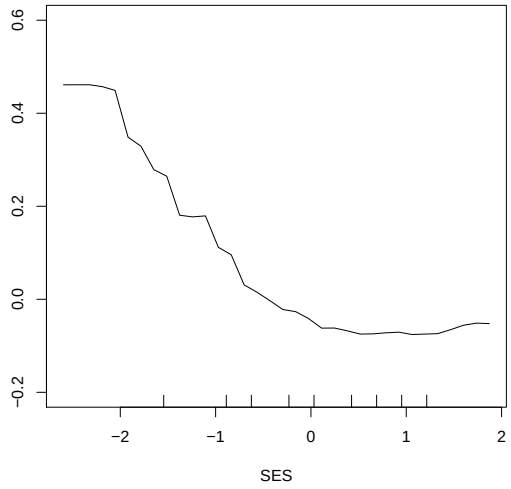


Figure 10:
Partial effect of the predictor *socio-economic status* (SES) (abscissa) on the dependent variable (ordinate) implied by a random forest of regression trees. The rugs on the abscissa indicate the deciles of the predictor variable.



5 Discussion

This paper briefly reviews the use of regression trees and random forests in psychological research and empirically illustrates the techniques employing data from the Erlangen-Nuremberg Development and Prevention Study. Regression trees and random forests enable the researcher to model the dependent variable without imposing a priori assumptions on the model structure such as specifying a particular functional form or choosing a set of predictors that affect the dependent variable. Since fitting regression trees is computationally cheap, the approach may serve as a preliminary benchmark for the considered modeling assumptions and may – in applications where we lack guidance from psychological theory or existing evidence – provide first hints concerning their validity for the practitioner. However, regression trees may be unstable and exhibit low predictive performance. Random forests do not suffer from these drawbacks as the technique basically averages over many regression trees, while fitting the model on bootstrap samples drawn from the original data (and, therefore, decorrelating the individual trees by reducing the overlap of the data based on which the trees are fit) and considering only a subset of the predictors to make a particular split (further decorrelating the individual trees). Employing random forests may be a valuable extension to the toolbox of the researcher, as the technique allows to validate existing results or theories in a flexible, data-driven way in the spirit of Breiman (2001b).

In the empirical application of the paper, we were able to replicate main results of Haupt et al. (2014) in a wider sense. Although the model structure and the choice of predictors varied from the data-driven random forest to the model fit based on a priori assumptions in Haupt et al. (2014), both models revealed similar facts: First, the in- and out-of-sample error measures computed for both models were comparable. Second, in agreement with the quantile regression approach in Haupt et al. (2014), we found different effects at different levels of the predictors (e.g., with regard to SES). Hence, both modeling approaches indicated that a conventional multiple linear regression may not be sufficient to capture the complex dependence structure underlying the conditional distribution of the dependent variable.

Although the main aims of our analysis were the description and demonstration of statistical methods, the models also showed interesting findings on the psychological topic of our application. The data-driven approach revealed a small number of risk factors that were relevant for fathers' corporal punishment of children. Physical punishment was most prevalent where the interaction between parents and children was dysfunctional and the fathers had enhanced levels of physical health complaints. Other somewhat less important risk factors were inconsistent discipline in parenting, a practice of various forms of discipline, and perceived difficulties of the child. The socio-economic circumstances of the family were also important, but their influence varied across the variable. Other demographic factors like age of parents were not relevant.

The most relevant risk factors of the training model could be cross-validated and therefore seem to be useful for practice. Although they were "detected" in a data-driven approach,

the key findings were consistent with hypotheses and findings on risk factors for corporal punishment and physical abuse in the international literature (Bender & Lösel, 2015; Gershoff, 2002; Peltonen et al., 2014). For example, parents who more often use corporal punishment for disciplinary purposes have low income, low educational achievement, and higher levels of mental stress (Bender & Lösel, 2015; Combs-Orme & Cain, 2008; de Paula Gebara et al., 2017). Fathers' personality also played a (minor) role, but this influence may depend on the cultural context (Masuda et al., 2019).

Overall, we conclude that aggressive parenting behavior often occurs when the families are under psychosocial stress, conflicts between parents and children escalate, and parents are somewhat helpless and inconsistent in their parenting behavior. These circumstances seem to be more relevant than basic personality dispositions or demographic variables (with the exception of low SES). Most of the above-mentioned results had rather small effect sizes. This underlines the general finding in psychological and criminological risk research that the accumulation of risks is important for validity (Lösel & Bender, 2006). The rather small effect sizes are insofar plausible as the data in our study stem from a "normal" community sample that did not contain many families at high risk for corporal punishment and physical child abuse. An oversampling of high-risk families would probably have led to more extreme cases and stronger correlations. In addition, the parental self-report on physical punishment may have been influenced by social desirability, i.e., to adhere to widely accepted norms of "appropriate" parenting. We do not know the amount of these influences, but assume that they may have reduced empirical effects in our study. On the other hand, our model tests only included one measurement point what may have led to an over-estimation of effects. It is necessary and we are planning to replicate the findings by longitudinal data.

References

Abidin, R. R. (1995). Parenting stress index (psi). *Odessa, FL: Psychological Assessment Resources (PAR)*.

Afifi, T. O., Ford, D., Gershoff, E. T., Merrick, M., Grogan-Kaylor, A., Ports, K. A., … Bennett, R. P. (2017). Spanking and adult mental health impairment: The case for the designation of spanking as an adverse childhood experience. *Child Abuse & Neglect*, *71*, 24–31.

Baumrind, D., Larzelere, R. E., & Cowan, P. A. (2002). Ordinary physical punishment: Is it harmful? Comment on Gershoff (2002). *Psychological Bulletin*, *128*(4), 580–589.

Bender, D., & Lösel, F. (2005). 20. Misshandlung von Kindern: Risikofaktoren und Schutzfaktoren [Abuse of children: Risk factors and protective factors]. In M. Herzog-Evans & D. I (Eds.), *Kindesmisshandlung und Vernachlässigung: Ein Handbuch [Child abuse and neglect: A handbook* (pp. 317–346). Hogrefe Verlag.

Bender, D., & Lösel, F. (2015). Risikofaktoren, Schutzfaktoren und Resilienz bei Misshandlung und Vernachlässigung [Risk factors, protective factors, and resilience in child abuse and neglect]. In U. T. Egle, P. Joraschky, A. Lampe, I. Seiffge-Krenke, & M. Cierpka (Eds.),

Sexueller Missbrauch, Misshandlung, Vernachlässigung [Sexual abuse, physical abuse, and neglect, 4th ed.] (pp. 77–103). Schattauer.

Bernard, T. J. (1990). Twenty years of testing theories: What have we learned and why? *Journal of Research in Crime and Delinquency*, *27*(4), 325–347.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Bruinsma, G. (2016). Proliferation of crime causation theories in an era of fragmentation: Reflections on the current state of criminological theory. *European Journal of Criminology*, *13*(6), 659–676.

Combs-Orme, T., & Cain, D. S. (2008). Predictors of mothers' use of spanking with their infants. *Child Abuse & Neglect*, *32*(6), 649–657.

de Paula Gebara, C. F., Ferri, C. P., de Castro Bhona, F. M., de Toledo Vieira, M., Lourenço, L. M., & Noto, A. R. (2017). Psychosocial factors associated with mother–child violence: A household survey. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 77–86.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192.

Ellonen, N., Peltonen, K., Pösö, T., & Janson, S. (2017). A multifaceted risk analysis of fathers' self-reported physical violence toward their children. *Aggressive Behavior*, *43*(4), 317–328.

Fahrenberg, J., Hampel, R., & Selg, H. (1989). Das Freiburger Persönlichkeitsinventar: revidierte Fassung FPI-R und teilweise geänderte Fassung FPI-A1 [The Freiburg Personality Inventory - revised version (FPI-R)]. Verlag für Psychologie, Hogrefe.

Ferguson, C. J. (2013). Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. *Clinical Psychology Review*, *33*(1), 196–208.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta (Ed.), *Machine learning: Proceedings of the thirteenth international conference on machine learning (icml '96)* (pp. 148–156). Bari, Italy.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407.

Geißler, R. (1994). *Soziale Schichtung und Lebenschancen in Deutschland (2., völlig neu bearbeitete und aktualisierte Auflage) [Social Stratification and Life Chances in Germany (Second Edition)]*. Stuttgart: Ferdinand Enke Verlag.

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, *128*(4), 539–579.

Gershoff, E. T. (2010). More harm than good: A summary of scientific research on the intended and unintended effects of corporal punishment on children. *Law & Contemporary Problems*, *73*, 31–56.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction (second edition)*. Springer.

Haupt, H., Lösel, F., & Stemmler, M. (2014). Quantile regression analysis and other alternatives to ordinary least squares regression: A methodological comparison on corporal punishment. *Methodology*, *10*(3), 81–91.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2019, March 6). *Party: A laboratory for recursive partytioning*. Version 1.3-3. Retrieved from https://cran.r-project.org/web/packages/party/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*, *14*, 3–15. doi:https://doi.org/10.1027/1614-2241/a000141

Kleinke, K., Stemmler, M., Reinecke, J., & Lösel, F. (2011). Efficient ways to impute incomplete panel data. *AStA Advances in Statistical Analysis*, *95*(4), 351–373.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.

Larzelere, R. E., & Baumrind, D. (2010). Are spanking injunctions scientifically supported. *Law & Contemporary Problems*, *73*, 57–87.

Liaw, A., & Wiener, M. (2018, March 25). *Randomforest: Breiman and cutler's random forests for classification and regression*. Version 4.6-14. Retrieved from https://cran.r-project.org/web/packages/randomForest/

Loeber, R., & Farrington, D. P. (1998). Never too early, never too late: Risk factors and successful interventions for serious and violent juvenile offenders. *Studies on Crime & Crime Prevention*, *7*(1), 7–30.

Lösel, F. (2017). Self-control as a theory of crime: A brief stocktaking after 27/42 years. In C. Bijleveld & P. van der Laan (Eds.), *Liber amicorum gerben bruinsma* (pp. 232–238). Boom Criminologie.

Lösel, F. (2018). Evidence comes by replication, but needs differentiation: The reproducibility issue in science and its relevance for criminology. *Journal of Experimental Criminology*, *14*(3), 257–278.

Lösel, F., & Bender, D. (2006). Risk factors for serious juvenile violence. In A. Hagell & R. Jeyarajah-Dent (Eds.), *Children who commit acts of serious interpersonal violence: Messages for best practice* (pp. 42–72). Jessica Kingsley Publishers London.

Lösel, F., Stemmler, M., & Bender, D. (2013). Long-term evaluation of a bimodal universal prevention program: Effects on antisocial development from kindergarten to adolescence. *Journal of Experimental Criminology*, *9*(4), 429–449.

Lösel, F., Stemmler, M., Jaursch, S., & Beelmann, A. (2009). Universal prevention of antisocial development: Short- and long-term effects of a child- and parent-oriented program. *Monatsschrift für Kriminologie und Strafrechtsreform [Monthly Journal for Criminology and Penal Law Reform]*, *92*(2-3), 289–307.

Masuda, R., Lanier, P., & Hashimoto, H. (2019). The association between paternal job stress and maternal child corporal punishment: Evidence from a population-based survey in metropolitan Japan. *Journal of Family Violence*, *34*(2), 119–126.

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods*, *21*(4), 583–602.

Monahan, J. (2012). The individual risk assessment of terrorism. *Psychology, Public Policy, and Law*, *18*(2), 167–205.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Peltonen, K., Ellonen, N., Pösö, T., & Lucas, S. (2014). Mothers' self-reported violence toward their children: A multifaceted risk analysis. *Child Abuse & Neglect*, *38*(12), 1923–1933.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Scarr, S., & Deater-Deckard, K. (1997). Family effects on individual differences in development. In S. S. Luthar, J. A. Burack, D. Cicchetti, & W. J. R (Eds.), *Developmental psychopathology: Perspectives on adjustment, risk, and disorder* (pp. 115–136). Cambridge University Press.

Seay, D. M., Jahromi, L. B., Umaña-Taylor, A. J., & Updegraff, K. A. (2016). Intergenerational transmission of maladaptive parenting strategies in families of adolescent mothers: Effects from grandmothers to young children. *Journal of Abnormal Child Psychology*, *44*(6), 1097–1109.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology*, *25*(3), 317–329.

Straus, M. A. (2009). *Beating the devil out of them: Corporal punishment in American families and its effects on Cildren (third printing)*. New Brunswick, NJ: Transaction Publishers.

Straus, M. A. (2010). Criminogenic effects of corporal punishment by parents. In M. Herzog-Evans (Ed.), *Transnational criminology manual* (pp. 373–390). Amsterdam: Wolf Legal Publishing.

Therneau, T., & Atkinson, B. (2019, April 10). *Rpart: Recursive partitioning and regression trees*. Version 4.1-15. Retrieved from https://cran.r-project.org/web/packages/rpart/

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wallner, S., Lösel, F., Stemmler, M., & Corrado, R. (2018). The validity of the cracow instrument in the prediction of antisocial development in preschool children: A five-year longitudinal community-based study. *International Journal of Forensic Mental Health*, *17*(2), 181–194.

Widom, C. S., Czaja, S. J., & DuMont, K. A. (2015). Intergenerational transmission of child abuse and neglect: Real or detection bias? *Science*, *347*(6229), 1480–1485.