Advances in Data Analysis and Classification (2023) 17:431–461
https://doi.org/10.1007/s11634-022-00509-3

REGULAR ARTICLE

Mixed-effect models with trees

Anna Gottard[1] · Giulia Vannucci[1] · Leonardo Grilli[1] · Carla Rampichini[1]

Received: 10 November 2021 / Revised: 30 May 2022 / Accepted: 6 June 2022 /
Published online: 8 July 2022

Abstract

Tree-based regression models are a class of statistical models for predicting continuous response variables when the shape of the regression function is unknown. They naturally take into account both non-linearities and interactions. However, they struggle with linear and quasi-linear effects and assume *iid* data. This article proposes two new algorithms for jointly estimating an interpretable predictive mixed-effect model with two components: a linear part, capturing the main effects, and a non-parametric component consisting of three trees for capturing non-linearities and interactions among individual-level predictors, among cluster-level predictors or cross-level. The first proposed algorithm focuses on prediction. The second one is an extension which implements a post-selection inference strategy to provide valid inference. The performance of the two algorithms is validated via Monte Carlo studies. An application on INVALSI data illustrates the potentiality of the proposed approach.

Keywords Multilevel data · Interaction effects · Recursive partitioning · Regression trunk models

Mathematics Subject Classification 62J99 · 62J05 · 62G08

1 Introduction

Mixed-effect or multilevel models (Snijders and Bosker 2012; Pinheiro and Bates 2006) are a valuable class of models able to deal with hierarchical/clustered data. Typical hierarchical data consist of statistical units (level 1 units) nested into clus-

The authors thank the Associate Editor and the anonymous reviewers for their very useful comments that led to this improved version of the paper.

✉ Anna Gottard
anna.gottard@unifi.it

[1] Department of Statistics, Computer Science, Applications "G. Parenti", Florence Center for Data Science, University of Florence, Florence, Italy

ters (level 2 units). Classic examples are students clustered within schools (individual cross-sectional data) or children's growth evaluated at several time points (repeated measures). Mixed-effect models consider unit clustering, including both fixed and random effects in the model. The standard linearity assumption for the fixed effects is too stringent in many situations, and a more flexible model specification, including non-linear effects and interactions, might be required. A worthwhile approach exploits regression trees (CART) (Breiman et al. 1984) to capture non-linear fixed effects and interactions via a piece-wise constant regression function. Several tree-based algorithms have been proposed in the literature to improve CART performance in different ways. For instance, see Loh (2002), Hothorn et al. (2006), Dusseldorp et al. (2010) and the subsequent literature to ensemble learning.

Tree-based models typically assume *iid* data and therefore ignore a clustered data structure when present. As it is well known that ignoring the clustering structure can lead to biased inference (see, e.g. Bryk and Raudenbush 2001, for linear models), some proposals have been suggested in the literature to adapt tree-based models to clustered, multilevel and longitudinal data. In the framework of longitudinal data, Segal (1992) was, up to our knowledge, the first to deal with this topic, generalizing the CART algorithm and its loss function to the case of correlated multiple binary responses. Zhang (1998) discussed the case of multiple binary response variables using as impurity measure the generalized entropy criterion, linked to the log-likelihood of a specific exponential family distribution. Other contributions are due to Abdolell et al. (2002), Loh and Zheng (2013), Eo and Cho (2014), among many. Being specific for longitudinal data, these solutions cannot be adopted when the clustered data do not have a regular structure.

In the general framework of multilevel data, regression trees have been extended to clustered data by Hajjem et al. (2011) and Sela and Simonoff (2012) modelling fixed effects with a decision tree while automatically accounting for random effects with a linear mixed model in a separate step. In particular, Hajjem et al. (2011) proposed mixed-effect regression trees (MERT) where CART selects the tree that models the fixed effect part, and a node-invariant linear structure is used to model the random effects. The algorithm is implemented within the framework of the expectation-maximization algorithm. Sela and Simonoff (2012) independently proposed a similar solution, called Random Effects/EM trees (RE-EM tree). It is shown that random effect regression trees are less sensitive to parametric assumptions and provide substantially improved predictive power compared to linear models with random effects and regression trees without random effects. The literature has grown with variants and extensions (e.g. Hajjem et al. 2014; Fu and Simonoff 2015; Hajjem et al. 2017; Miller et al. 2017; Pellagatti et al. 2021). Other solutions, including trees and clustered data, have been proposed in the framework of subgroups analysis. See for instance Fokkema et al. (2018) and Seibold et al. (2019).

This paper proposes a semi-parametric mixed-effect model where trees are added to a linear component for capturing interactions and non-linear effects. The idea of adding a linear component to a tree is not new in the literature. While regression trees can easily capture complex dependencies by reconstructing the entire regression function, they need many fortuitous splits to recreate a linear or quasi-linear function (Friedman et al. 2001). For this reason, tree-based models are said to struggle in fitting linear or additive dependencies. To overcome this issue, Dusseldorp and Meulman (2004) were the first to propose to combine a linear part and a regression tree in a single model, called regression trunk model, that can be efficiently estimated by a Simultaneous Threshold Interaction Modelling Algorithm (STIMA) proposed by Dusseldorp et al. (2010). Here we extend regression trunk models to the case of the mixed-effect models. Moreover, we include multiple trees in the tree component instead of one. For the particular application we have in mind, we consider here the case of three trees. The first tree captures non-linear relationships and interactions among level 1 predictors. The second tree captures non-linear relationships and interactions among level 2 predictors, while the third is devoted to detecting cross-level interactions. This last kind of interaction is of particular interest in multilevel analysis (see, for instance, Bauer and Curran 2005).

Following the reasoning of Efron (2020), the proposed approach lies in between a pure predictive algorithm and a more traditional regression method, providing an easy to interpret mixed-effect model with good predictive properties. The linear component helps to maintain the three trees as short as possible, while the three trees act as *weak learners* from the point of view of prediction. The resulting model is more easily interpretable than a single tree or a random forest. Moreover, it has a better predictive performance than a linear mixed-effect model. We call this class *Three-tree mixed-effect* (3Trees) models.

For the estimation of 3Trees models, we propose two different algorithms. The first algorithm relies on a backfitting-like procedure for selecting the three trees component, and then it jointly estimates the linear and tree-based parts. It provides good predictions of the response Y for new units in already observed or new clusters. Therefore, this algorithm is useful when a study has predictive purposes only. The second algorithm modifies the first one by introducing a post-selection inference procedure, which provides a pruning method based on hypothesis tests and valid confidence intervals for the predicted values. Specifically, we apply a split sample procedure (Cox 1975) on clusters. We evaluate the proposed algorithms through simulations. It turns out that both algorithms are computationally efficient and have satisfactory performance.

The remainder of the paper is organized as follows. Section 2 describes the proposed Three-tree mixed-effect model and its interpretation. Section 3 presents the two estimating algorithms. Section 4 reports a simulation study to evaluate the performance of the two algorithms in the case of a random intercept model. Section 5 illustrates an application of the 3Trees model to analyze the Maths scores achieved by Italian children at a national standardized test. The last section offers concluding remarks and outlines directions for future work.

2 The Three-tree mixed-effect model

In this section, we present the mixed-effect model called *Three-tree mixed-effect model* (3Trees model). For simplicity, we focus on the case of random intercept mixed models, but the extension to random slopes is straightforward. See Appendix B for an example.

Suppose that Y_{ij} is a quantitative response variable measured on unit i, $i = 1, \dots, n_j$, belonging to cluster/group j, with $j = 1 \dots, J$ and $n_{tot} = \sum_{j=1}^{J} n_j$.

We assume that the *true* data generating process has the form

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_j\boldsymbol{\gamma} + g(\mathbf{x}_{ij}, \mathbf{z}_j) + u_j + \varepsilon_{ij} \tag{1}$$

where $\mathbf{x}_{ij}$ is the vector of p_1 individual-level predictors, $\boldsymbol{\beta}$ the associated main fixed effect coefficients, including β_0 as intercept, while $\mathbf{z}_j$ is the vector of p_2 cluster-level predictors and $\boldsymbol{\gamma}$ the associated main fixed effect coefficients. Let $p = p_1 + p_2$. Here $g(\cdot)$ is an unknown function $\mathbb{R}^p \to \mathbb{R}$, ruling the non-linear fixed-effects, assumed to not vary across clusters. The level 1 errors ε_{ij} are assumed to be *iid* $N(0, \sigma^2_\varepsilon)$. The level 2 errors (random effects) u_j are supposed to be *iid* $N(0, \sigma^2_u)$. Moreover, the level 1 and level 2 errors are assumed to be independent. Thus, the model assumes that responses of units belonging to the same cluster are positively correlated, with correlation equal to

$$ICC = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon}, \tag{2}$$

although responses of units belonging to different clusters are uncorrelated. In model (1), for each cluster j, the random quantity $b_j = \beta_0 + u_j$ is the so-called random intercept, varying across clusters, with $b_j \perp\!\!\!\perp b_{j'}$ for all $j \neq j' \in \{1, \ldots, J\}$.

As the non-linear function $g(\cdot)$ is unknown, we propose a nonparametric approximation based on regression trees. In particular, considering a two-level structure, we propose to use three trees. The first tree accounts for non-linearities at level 1, the second tree accounts for non-linearities at level 2, and the third one accounts for non-linearities among all predictors. The primary purpose of the third tree is to account for cross-level interactions, namely interactions among level 1 and level 2 predictors. We call this model the Three-tree mixed-effect (3Trees) model

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_j\boldsymbol{\gamma} + \sum_{t=1}^{3} T_t(\mathbf{h}_{tij}) + u_j + \varepsilon_{ij}. \tag{3}$$

In the 3Trees model, the function $g(\cdot)$ of Eq. (1) is approximated by the sum of three regression trees, T_t ($t = 1, 2, 3$), each constructed over a subset of predictors. Specifically, $\mathbf{h_1}_{ij} \subseteq \mathbf{x}_{ij}$ (level 1 predictors), $\mathbf{h_2}_{ij} \subseteq \mathbf{z}_j$ (level 2 predictors) and $\mathbf{h_3}_{ij} \subseteq (\mathbf{x}_{ij} \cup \mathbf{z}_{ij})$ (level 1 and level 2 predictors).

The model is additive in its components, and the tree component stands as a region-specific intercept. Model (3) can be written equivalently as

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_j\boldsymbol{\gamma} + \sum_{t=1}^{3}\sum_{m=1}^{M_t} \mu_{tm}\mathbb{I}\{\mathbf{h}_{tij} \in R_{tm}\} + u_j + \varepsilon_{ij}, \tag{4}$$

where $R_{t1}, \ldots, R_{tM_t}$ is the partition of the predictor space corresponding to the tree T_t. From this point of view, each tree acts as a factor with M_t categories. Each category corresponds to a tree leaf, representing a region of the selected partition of the predictor space. The region R_{tm} can be identified by multiplying all the dummy variables defined by the binary splits along the path from the root node to each leaf in the tree T_t. Therefore, some classical identifiability constraints, such as sum-to-zero or corner constraints, must be included for each of these three factors.

When the unknown regression function can be assumed to be quasi-linear (Wermuth and Cox 1998), the number of leaf nodes M_t $(t = 1, 2, 3)$ can be kept small. In this situation, the proposed class of models is most convenient, providing a good balance between performance and interpretability. The presence of three trees instead of one, e.g. as used for *iid* data in STIMA (Dusseldorp et al. 2010), allows us to achieve the same goodness of fit with trees having smaller depth. In addition, devoting the trees to specific sets of predictors improves the ability of the algorithm to detect different kinds of non-linear effects.

The choice of the subsets of variables $\mathbf{h_1}_{ij}$ $\mathbf{h_2}_{ij}$ and $\mathbf{h_3}_{ij}$ to be included in each tree is driven by subject-matter considerations. As a general practice, we recommend including all the potentially relevant variables and leaving it to the algorithm to select the most pertinent. Notice that the presence of the linear component attenuates the typical issue of the tree-based algorithms with correlated predictors.

It is worth noting that if the regression function has strong non-linearities, with abrupt inversions in the sign of dependence, model (3) may require deep trees to achieve good performance. One can include flexible functions in the linear part to overcome this issue, such as polynomials or splines. In this case, the so-called linear component captures such strong non-linearities, while the tree component mostly picks up the interaction effects.

The main difference of our procedure with respect to previous proposals (Hajjem et al. 2011; Sela and Simonoff 2012), is the inclusion of the linear component $\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_j\boldsymbol{\gamma}$ in the mixed-effect model (3). This inclusion allows to avoid overfitting and helps interpretation.

3 Two estimating algorithms

We propose two iterative procedures to select the trees and estimate the model parameters. While previous proposals for trees with clustered data were based on a kind of EM algorithm (see, for instance, Hajjem et al. 2011), we propose a procedure similar to backfitting for the selection of the trees.

Recalling that we denoted the total number of sample units as $n_{tot} = \sum_{j=1}^{J} n_j$, let us call $\boldsymbol{X}$ the $n_{tot} \times p_1$ matrix of the individual-level predictors, with rows $\mathbf{x}'_{ij}$. Let $\boldsymbol{Z}$ be the $n_{tot} \times p_2$ matrix of the cluster-level predictors, with rows $\mathbf{z}_{ij} = \mathbf{z}'_j$ for all the units of the same cluster, $i = 1, \ldots, n_j$. Denote $\boldsymbol{H}_L = (\boldsymbol{X}, \boldsymbol{Z})$ the $n_{tot} \times p$ matrix, with $p = p_1 + p_2$ of all the predictors included into the linear component, $\boldsymbol{H}_1$ the matrix $n_{tot} \times h_1$, $h_1 \leq p_1$, whose columns are selected from $\boldsymbol{X}$ to be considered for the tree T_1 , $\boldsymbol{H}_2$ the matrix $n_{tot} \times h_2$, $h_2 \leq p_2$, whose columns are selected from $\boldsymbol{Z}$ to be considered for the tree T_2 and $\boldsymbol{H}_3$ the matrix $n_{tot} \times h_3$, $h_3 \leq p$, whose columns are selected from $\boldsymbol{H}_L$ for T_3. Finally, call $\boldsymbol{y}$ the $n_{tot} \times 1$ vector of the responses.

The first algorithm we are proposing, called 3Trees-Alg and whose pseudo-code can be found in Algorithm 1, has predictive purposes. It consists of two main steps:

Algorithm 1: 3Trees-Alg Backfitting algorithm for Three-tree linear mixed models

```
Input: $\mathcal{D} = \{y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_j\}$ , $i = 1, \dots, n_j$ , $j = 1, \dots, J$ and the derived matrices
       $\boldsymbol{H}_L, \boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3$
Output: Response predictions based on the tree embedded linear mixed model (4)
Initialization: $\widehat{\boldsymbol{y}}_{T1}^{(0)} = \widehat{\boldsymbol{y}}_{T2}^{(0)} = \widehat{\boldsymbol{y}}_{T3}^{(0)} = \overline{y}/3$ ; $\widehat{\boldsymbol{y}}_L^{(0)} = (\boldsymbol{H}'_L\, \boldsymbol{H}_L)^{-1}\, \boldsymbol{H}'_L\, \boldsymbol{y}$ ;
Selection step. Select model $\widehat{\mathcal{M}}$ with $\min_l$ MSE$^{(l)}$:
repeat
    sub-step L:
    Compute the vector of partial residuals for the linear component as $\boldsymbol{r}_{\mathrm{L}}^{(l)} = \boldsymbol{y} - \sum_{t=1}^{3} \widehat{\boldsymbol{y}}_{Tt}^{(l-1)}$;
    Fit a linear mixed-effect model of $\boldsymbol{r}_{\mathrm{L}}^{(l)}$ on $\boldsymbol{H}_L$;
    Predict $\widehat{\boldsymbol{y}}_L^{(l)}$ from the fitted model using both fixed and random effects

    sub-step Tree1:
    Compute the vector of partial residuals for $T_1$ as $\boldsymbol{r}_1^{(l)} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_L^{(l)} - \widehat{\boldsymbol{y}}_{T2}^{(l-1)} - \widehat{\boldsymbol{y}}_{T3}^{(l-1)}$;
    Fit a regression tree with the CART algorithm for $\boldsymbol{r}_1^{(l)}$ on $\boldsymbol{H}_1$;
    Predict $\widehat{\boldsymbol{y}}_{T1}^{(l)}$ from the fitted tree

    sub-step Tree2:
    Compute the vector of partial residuals for $T_2$ as $\boldsymbol{r}_2^{(l)} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_L^{(l)} - \widehat{\boldsymbol{y}}_{T1}^{(l)} - \widehat{\boldsymbol{y}}_{T3}^{(l-1)}$;
    Fit a regression tree with the CART algorithm of $\boldsymbol{r}_2^{(l)}$ on $\boldsymbol{H}_2$;
    Predict $\widehat{\boldsymbol{y}}_{T2}^{(l)}$ from the fitted tree

    sub-step Tree3:
    Compute the vector of partial residuals for $T_3$ as $\boldsymbol{r}_3^{(l)} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_L^{(l)} - \widehat{\boldsymbol{y}}_{T1}^{(l)} - \widehat{\boldsymbol{y}}_{T2}^{(l)}$ ;
    Fit a regression tree with the CART algorithm of $\boldsymbol{r}_3^{(l)}$ on $\boldsymbol{H}_3$ ;
    Predict $\widehat{\boldsymbol{y}}_{T3}^{(l)}$ from the fitted tree

    sub-step evaluation: Compute MSE$^{(l)}$ = mean$\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_L^{(l)} - \widehat{\boldsymbol{y}}_{T1}^{(l)} - \widehat{\boldsymbol{y}}_{T2}^{(l)} - \widehat{\boldsymbol{y}}_{T3}^{(l)}\right)^2$
until convergence criterion is met;
Estimation step: Compute parameter estimates for model $\widehat{\mathcal{M}}$ using the standard maximum
    likelihood estimator for linear mixed-effect model of $y$ on $\boldsymbol{H}_L$ and the three factors corresponding to
    the trees selected at iteration $l$ : MSE = argmin MSE$^{(l)}$.
```

the *Selection step* and the *Estimation step*. The Selection step aims to find the best trees conditionally on the linear component. The algorithm initialises the sum of the three trees equal to the sample mean and the linear component at the standard least squared estimates. This choice provides a neutral starting point that, in our experience, facilitates algorithm convergence. Then the procedure iteratively runs the CART algorithm (Breiman et al. 1984) on the partial residuals with respect to the other trees and the mixed effect linear component. At the end of each iteration, the algorithm computes the Mean Squared Error (MSE) based on the predictions for each tree and the mixed-effect model. The iterative procedure stops when the *convergence criterion* is met, namely, the difference between MSE in two successive iterations is below a given threshold, or the maximum number of iterations is achieved. Notice that as the regression function is not smooth, the algorithm can fall into a periodic loop. Hiabu et al. (2021), in a different framework, proved the validity of backfitting in the presence of non-smoothness. Therefore, the Selection step of 3Trees-Alg returns the trees corresponding to the minimum MSE. For the Selection step, one has also to set the tuning parameters of the CART algorithm, such as the maximum depth of the tree, the minimum number of units in a leaf node and the complexity parameter CART. If this latter parameter is set to a value different from zero, then the CART algorithm within 3Trees-Alg performs tree pruning at each iteration via cross-validation.

In the second step of the 3Trees-Alg algorithm, the Estimation step, each tree selected in the Selection step is specified as a factor, and a mixed-effect model with the linear component and the three new factors as in (4) is fitted via maximum likelihood. The step returns the estimates of all the parameters μ_{tm}, with $m = 1, \ldots, M_t$, $t = 1, 2, 3$, for the tree component, the vectors of coefficients for the linear components $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the variances $\sigma_u^2, \sigma_\varepsilon^2$.

It is worth pointing out that the tree component is selected non-parametrically. The CART procedure to choose the trees is greedy, and therefore it addresses the splittings of the predictors providing the local largest decrease in MSE. This implies a lack of order invariance in the tree component. Namely, if we permute the order of the trees in the Selection step, the resulting selected trees can be different. The order invariance can be guaranteed when the subsets of variables included in the trees are disjoint, i.e. each variable can participate in the construction of only one tree, and the variables are independent between trees. However, in multilevel settings, we are often interested in detecting cross-level interactions, which requires defining a third tree that includes all predictors. Thus the trees cannot be kept disjoint. In principle, one could try all six possible orderings of trees and choose by predictive performance comparison. However, we suggest the order dictated by the standard strategy of model building in multilevel analysis (Snijders and Bosker 2012), namely first select level 1 predictors, then level 2 predictors, and finally cross-level interactions.

The 3Trees-Alg algorithm, described in Algorithm 1, is implemented in a user written R code (avalaible from the authors), using the package `lme4` (Bates et al. 2015a) for the estimation of the mixed models and on `rpart` (Therneau and Atkinson 2019) for selecting the trees via the CART algorithm.

3.1 Algorithm based on post-selection inference

The proposed tree-based mixed-effect model inherits the theoretical properties of parametric models if the model is well specified. If the selected regression function is not assumed to be the *true* one but rather a good approximation of it, inference is referred to the projection parameters (see, for instance Buja et al. 2019), where the regression function is optimally projected into the space of the semilinear functions in (4). The tree component identification utilizes the CART algorithm that is proved to provide a locally optimal solution (Breiman et al. 1984).

In a context where the model is selected using the data, as in our procedure, classical tools for inference, such as p-values and confidence intervals, are invalid (see, for instance Benjamini 2010; Berk et al. 2013). In particular, as confidence intervals are computed after model selection, the classical procedures provide narrower confidence intervals than due at the nominal level, with an actual coverage below the nominal one. Interesting literature on post-selection inference proposes methods mainly in the context of ordinary least square and Lasso-type estimators for *iid* data. A simple post-selection inference procedure, proper when the interest is to discover dependencies and influential predictors in a data-driven model, is the *split sampling* procedure (Cox 1975). It involves the random division of the sample into two parts, one used to select the model and one used for inference. This procedure allows deriving conditional (post-selection) tests or confidence intervals. In addition, when required, inference based on sample splitting followed by bootstrap provides assumption-lean, robust confidence intervals for the projection parameters (Rinaldo et al. 2019).

To obtain adequate inference for the parameters of a 3Trees model, we propose adopting a split sampling procedure. At this aim, we need to assume that the maximum size of the selected model is under control. This corresponds to fixing the maximum depth of each tree.

For multilevel data, the standard assumption of *iid* data applies at the cluster-level. Therefore, sample splitting is obtained by randomly partitioning the J clusters into two subsets, approximately of equal size. This partition induces a partition of the original data set $\mathcal{D}$ into two subsets, say $\mathcal{D}_S$ and $\mathcal{D}_E$. The first subset, $\mathcal{D}_S$, is used to select the three trees using the Selection step of Algorithm 1. The second subset, $\mathcal{D}_E$, is used for inference on model parameters using the standard linear mixed effect procedures and software. The proposed algorithm, called 3Trees-Alg-PSI, is summarized in Algorithm 2.

Algorithm 2: 3Trees-Alg-PSI Backfitting algorithm for the Three-tree linear mixed models using sample splitting procedure

```
Input: $\mathcal{D} = \{y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_j\}$ , $i = 1, \dots, n_j$ , $j = 1, \dots, J$ and the derived matrices
       $\boldsymbol{H}_L, \boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3$
Output: Estimates of Model (4) parameters with their asymptotic standard errors
1 Random splitting: Let $[J] := \{1, \dots, J\}$. Randomly choose the subset $[J_S] \subset [J]$, of cardinality
  $\approx \frac{J}{2}$, and define $[J_E] := [J] \setminus [J_S]$.
2 Selection step: Select $\widehat{\mathcal{M}}$ applying the sub–algorithm Selection step of Algorithm 1 to the subset of
  data $\mathcal{D}_S = \{y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_j\}$ , $i = 1, \dots, n_j$ , $j \in [J_S]$.
3 Estimation step: Compute parameter estimates for model $\widehat{\mathcal{M}}$ using the standard maximum
  likelihood estimator for mixed effect models on the subset of data $\mathcal{D}_E = \{y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_j\}$
  $i = 1, \dots, n_j$ , $j \in [J_E]$.
```

Applying 3Trees-Alg-PSI, we obtain an interpretable predictive model laying in between exploratory and confirmatory data analysis. The confidence intervals computed using 3Trees-Alg-PSI can also be used to prune a tree when the complexity parameter in the CART sub-step is set to zero or kept very small. For this purpose, consider the two regions R_{tm} and $R_{tm'}$ corresponding to the left and right terminal leaves, produced by the same variable splitting of the tree T_t. Let μ_{tm} and $\mu_{tm'}$ be the corresponding coefficients in the 3Trees model as in (4). Pruning the tree corresponds to joining R_{tm} and $R_{tm'}$. Therefore, the tree can be safely pruned whenever the null hypothesis

$$H_0 : \mu_{tm} = \mu_{tm'} \quad \text{vs} \quad H_A : \mu_{tm} \neq \mu_{tm'}$$

cannot be rejected or, equivalently, when the confidence intervals for the two parameters overlap. To perform a deeper pruning, one has to check the equality of all the coefficients involved in the pruning. The pruned 3Trees model is nested in the larger 3Trees model by imposing equality constraints on the parameters. Therefore, standard techniques for nested model parameters, such as the Likelihood Ratio test, can be safely utilized.

4 Simulation studies

In this section, we propose two simulation studies to evaluate the algorithms presented in Sect. 3 for fitting 3Trees models.

The first simulation study aims at evaluating the predictive performance of Algorithm 1 (3Trees-Alg). In contrast, the second one is designed to assess the inferential accuracy in terms of confidence intervals of Algorithm 2 (3Trees-Alg-PSI).

We consider three scenarios with different data generating processes. Each scenario includes two individual-level predictors, X_1 and X_2, having bivariate Normal distribution, $N_2(\mathbf{0}, \boldsymbol{\Sigma})$, and two cluster-level predictors, Z_1 and Z_2, with distribution $N_2(\mathbf{0}, \boldsymbol{\Gamma})$. We set the following values for the covariance matrices

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma} = \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{pmatrix}$$

The random components are $u_j \sim N(0, \sigma_u^2)$, with $\sigma_u^2 = 3$, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon^2 = 1$. Those values imply a high Intraclass Correlation Coefficient ($ICC = 0.75$), which allows us to clearly show the relevance of using the random effects in predicting the response. The data generating processes of the three scenarios differ in the shape of the regression function, as described in the following.

Scenario 1: we assume the following linear model

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + u_j + \varepsilon_{ij}$$

where $\beta_0 = 5$, $\beta_1 = 1$, $\beta_2 = 0$, $\gamma_1 = 1$ and $\gamma_2 = 0$. In this scenario, the true model is a linear mixed effect model.

Scenario 2: we assume a quasi-linear model that includes linear terms, threshold non-linear terms and interaction terms, as follows

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \mu_1 \mathbb{I}(X_{1ij} \geq 0) + \\ + \mu_2 \mathbb{I}(Z_{2j} < 0) + \mu_3 \mathbb{I}(X_{1ij} < 0) \mathbb{I}(Z_{2j} \geq 0) + u_j + \varepsilon_{ij}, \end{aligned}$$

where $\beta_0 = 5$, $\beta_1 = 1$, $\beta_2 = 0$, $\gamma_1 = 1$ and $\gamma_2 = 0$, $\mu_1 = 2$, $\mu_2 = 3$ and $\mu_3 = -3$.

Scenario 3: we use the non-linear model of Hajjem et al. (2014), except that we replace the third and fourth level 1 predictors (X_3 and X_4) with level 2 predictors, denoted as Z_1 and Z_2.

$$\begin{aligned} Y_{ij} = &\beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \mu_1 X_{2ij}^2 \\ &+ \mu_2 Z_{1j} \ln |X_{1ij}| + u_j + \varepsilon_{ij}, \end{aligned}$$

with $\beta_0 = 0\,,\ \beta_1 = 2\,,\ \beta_2 = 0\,,\ \gamma_1 = 4$ and $\gamma_2 = 0\,,\ \mu_1 = 2\,,\ \mu_2 = 2$.
In each scenario, some parameters of the linear part are set to zero to provide a glimpse of the ability of the proposed methods to correctly identify predictors with no linear effects in this small dimensional setting.

4.1 Predictive performance

For each scenario described in the previous section, we generate 500 data sets with $J = 500$ clusters of $n_j = 100$ units, for a total sample size of 50 000 units. We randomly split each cluster into two equal parts to form the train and the test sets, each composed of $J = 500$ clusters of $n_j = 50$ units.

We compare the predictive performance of Algorithm 1 (3Trees-Alg) for fitting the three-tree models with the following commonly used methods: RE-EM trees (Sela and Simonoff 2012), regression trees (CART) (Breiman et al. 1984), and a linear mixed-effect model with main effects only (Linear-ME) fitted with maximum likelihood. In addition, we compare the performance of mixed effect random forest as proposed by Hajjem et al. (2014). As a benchmark, we also report the results for a mixed-effect model specified as the data generating model (True-ME). Notice that in Scenario 1, the True-ME model and the Linear-ME model coincide.

We implement the 3Trees-Alg in R (R Core Team 2020), using the packages `rpart` (Therneau and Atkinson 2019) for the tree component and `lme4` (Bates et al. 2015b) for the mixed-effect linear component. As tuning parameters for the tree component we set `rpart maxdepth` tuning parameters at 2 for the first two scenarions, and at 3 for the third one. The `rpart` complexity parameter `cp` is set at 0.0001. The `rpart` package has also been used for CART, while `lme4` has also been used to fit linear mixed-effect models Linear-ME and True-ME. The RE-EM algorithm has been implemented using the R package `REEMtree` (Sela and Simonoff 2021) with the complexity parameter for pruning at 0.01 and the number of standard errors used in pruning set at 1. For the mixed-effect random forest, we used two different R functions. The first function, `MixRF`, that we used with the default setting, is implemented in the package `MixRF` (Wang et al. 2016). The second function, `MERF`, is implemented in the package `LongituRF` (Capitaine et al. 2021). We used the default setting for this function, but without the stochastic longitudinal component, not adequate for the scenarios considered in this section.

In clustered data, we can distinguish two types of predicted values for the response variable: (i) the value of the response for a statistical unit of a hypothetical (new) cluster and (ii) the value of the response for a statistical unit belonging to a cluster already included in the sample. The two types of prediction differ in the value assigned

Table 1 Monte Carlo averages and standard deviations of the Mean Squared Error and Predictive Mean Squared Error, computed with fixed effects only (MSE, PMSE), or including BLUP predictions of the random effects ($clus$MSE and $clus$PMSE). Scenarios 1, 2 and 3 ($J = 500$, $n_j = 50$ for both the training set and the test set, number of Monte Carlo runs: 500, except for MixRF and MERF with 250 runs)

Algorithm	$clus$MSE	MSE	$clus$PMSE	PMSE
	SCENARIO 1			
True-ME	0.980 (0.009)	3.992 (0.194)	1.020 (0.009)	3.991 (0.197)
3Trees	0.979 (0.009)	3.991 (0.194)	1.021 (0.010)	3.992 (0.197)
RE-EM	1.102 (0.033)	4.319 (0.263)	1.159 (0.035)	4.333 (0.264)
MixRF	0.307 (0.004)	3.224 (0.199)	1.085 (0.010)	3.998 (0.196)
MERF	0.238 (0.005)	2.768 (0.210)	1.106 (0.010)	3.669 (0.202)
CART		4.276 (0.205)		4.307 (0.210)
	SCENARIO 2			
True-ME	0.980 (0.009)	3.985 (0.195)	1.020 (0.009)	3.985 (0.198)
3Trees	0.981 (0.019)	3.982 (0.197)	1.022 (0.020)	3.983 (0.199)
RE-EM	1.330 (0.031)	5.113 (0.481)	1.386 (0.032)	5.115 (0.482)
Linear-ME	2.625 (0.046)	7.502 (0.314)	2.730 (0.047)	7.500 (0.314)
MixRF	0.312 (0.004)	3.367 (0.198)	1.094 (0.010)	4.156 (0.196)
MERF	0.241 (0.004)	2.744 (0.205)	1.113 (0.010)	3.656 (0.197)
CART		4.694 (0.259)		4.704 (0.262)
	SCENARIO 3			
True-ME	0.980 (0.009)	3.992 (0.194)	1.020 (0.009)	3.991 (0.197)
3Trees	6.079 (0.347)	9.156 (0.414)	6.421 (0.367)	9.262 (0.423)
3Trees (depth 6)	3.244 (1.048)	6.133 (1.057)	3.504 (1.103)	6.277 (1.078)
RE-EM	6.853 (0.416)	12.577 (0.729)	7.289 (0.433)	12.762 (0.726)
Linear-ME	13.690 (0.412)	16.931 (0.470)	14.200 (0.417)	16.931 (0.470)
MixRF	0.708 (0.029)	4.702 (0.297)	2.192 (0.087)	6.260 (0.307)
MERF	0.312 (0.011)	3.298 (0.285)	1.473 (0.048)	4.486 (0.270)
CART		11.298 (0.457)		11.551 (0.469)

to the random effect. In the first type, the random effect is set at its expectation, namely zero, thus predicting the response with the fixed part only. On the other hand, in the second type, the random effect is set at the BLUP prediction for the observed cluster, thus predicting the response with the fixed part plus the BLUP prediction (Robinson 1991; Skrondal and Rabe-Hesketh 2009).

To evaluate the predictive performance of the methods under comparison, we compute the Mean Squared Error on the train data and the Predictive Mean Squared Error on the test data (new data). We denote with MSE and PMSE, respectively, the measures referred to the predictions with the fixed part only. In addition, we denote with $clus$MSE and $clus$PMSE the measures when the prediction refers to a new unit in an observed cluster, i.e. including the random effect.

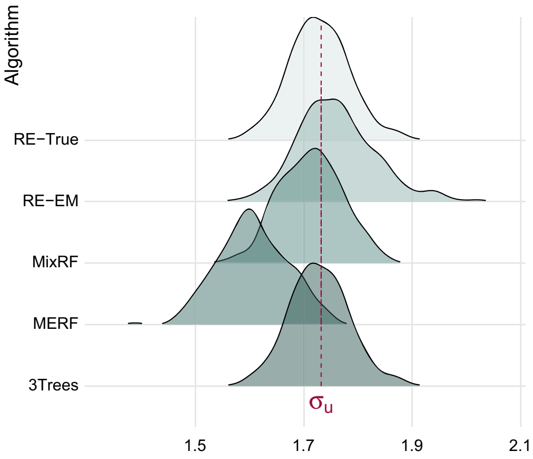
Table 1 reports the Monte Carlo averages and standard deviations of the mean squared errors for the methods under comparison. The predictive performance of the proposed 3Trees-Alg algorithm, as measured by *clus*PMSE and PMSE, is very close to the benchmark of the correctly specified mixed-effect model (True-ME) in Scenario 1 and Scenario 2. As expected, in the highly non-linear model of Scenario 3, the performance of 3Trees-Alg does not reach the benchmark. However, it is still better than the other algorithms, apart from the two mixed-effect random forest algorithms. It is worth noting that the performance of the 3Trees-Alg improves when the depth of the trees increases. As an example, for Scenario 3, we report the performance of 3Trees-Alg with trees of depth 6. This performance is substantially improved but at the cost of a less easy interpretation. An interesting aspect is the comparison of MSE and PMSE, which are pretty similar for the 3Trees algorithm, suggesting that it is never overfitting.

As expected, the error measures for predicting a unit in a new cluster (MSE and PMSE) are higher than the corresponding measures relative to the prediction for a unit of an observed cluster (*clus*MSE and *clus*PMSE). The magnitude is much larger (about three times) due to the high proportion of the between-cluster variation in the simulated data (ICC=0.75). It is worth noting that *clus*MSE and *clus*PMSE are obtained using BLUP predictions of level 2 errors, which are crucial for obtaining an accurate prediction for a unit of an observed cluster.

The standard deviation of the random effects σ_u is interesting *per se* and, in addition, it plays a key role in the prediction for an observed cluster since it affects the BLUP of the random effect. We compare the performance of the competing methods in estimating the parameter σ_u by plotting the Monte Carlo distributions (except for CART, which does not provide an estimate of σ_u). In Scenario 1 (Fig. 1), there are no relevant differences, except for MERF, which seems to underestimate the random effect variance.

On the other hand, Scenarios 2 and 3 are noteworthy, and the results are reported in Fig. 2. In all cases, the proposed method (3Trees-Alg) yields estimates of σ_u similar to the benchmark (True-ME), whereas the performance of RE-EM is not satisfactory. Indeed, RE-EM and MixRF tend to overestimate σ_u, with a large bias in Scenario 3,

Fig. 1 Monte Carlo distribution of the estimate of the parameter for the random effects standard deviation σ_u for the 3Trees and RE-EM algorithms, compared with the mixed-effect model with the *true* (linear) regression functions, under Scenario 1



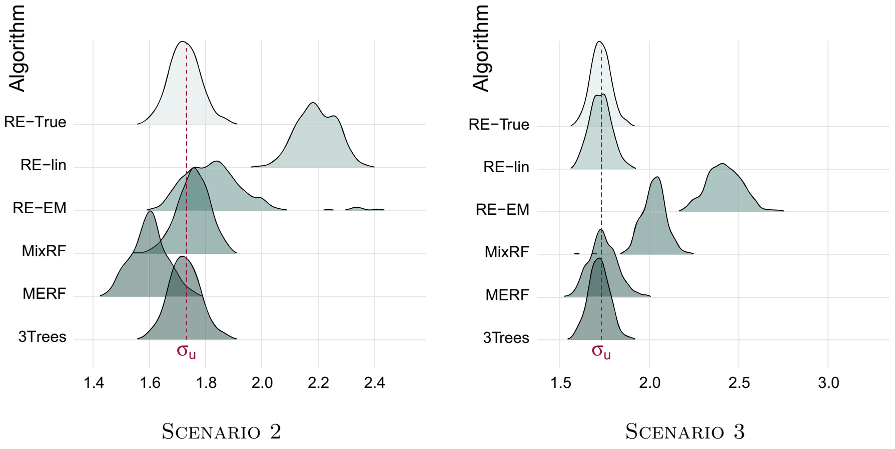


Fig. 2 Monte Carlo distributions of the estimate the random effects standard deviation σ_u for the 3Trees and RE-EM algorithms, compared with the mixed-effect models with the *true* and the linear regression functions respectively, under Scenario 2 and 3

while MERF underestimates it in Scenario 2. The 3Trees algorithm performs better than RE-EM and Linear-ME also in estimating the level 1 standard deviation σ_ε, as shown by the plots of the Monte Carlo distributions reported in Appendix A (Figs. 6 and 7).

Another aspect to consider is the estimate of the parameters of the linear component, β_1 , β_2 , γ_1 and γ_2. Notice that the proposed algorithm is not meant to recover the true data generating process but to predict the response by selecting a quasi-linear model closest to the true data generating process. When the tree component adequately approximates interactions and non-linearities, the parameters of the linear part should be close to the true ones. Table 2 reports the Monte Carlo averages and standard deviations of such estimates for the linear component. In both Scenario 1 and 2, the algorithm 3Trees-Alg provides estimates as good as the benchmark algorithm for null and non-null parameters. In Scenario 3, the parameters for the main effects of the level 1 predictors are correctly captured by the 3Trees-Alg algorithm, together with the null coefficient of the cluster-level predictor Z_2. Conversely, the coefficient of Z_1 turns out to be underestimated. This behaviour is due to the particular shape of the non-linear component, i.e. the interaction term $Z_{1j} \ln |X_{1ij}|$, which is probably inaccurately accounted for by the tree component. This behaviour confirms that this scenario would probably require a deeper tree to capture such a particular non-linear shape.

The simulation results are similar when we consider a smaller number of clusters, and smaller clusters ($J = 50$ and $n_j = 15$), with a test set for prediction evaluation of equal size. For these simulations, we also consider the predictive performance of a single regression trunk model for iid data (Dusseldorp et al., 2010) estimated via the R package `stima` with `maxsplit` set to 2. Table 7 of Appendix A reports the results for this case for the three scenarios. Simulations indicate that the behaviour of

Table 2 Monte Carlo averages and standard deviations of the estimates of the parameters in the linear component in Scenarios 1, 2 and 3 ($J = 500$, $n_j = 50$ for both the training set and the test set, number of Monte Carlo runs: 500)

Algorithm	Parameters			
	β_1	β_2	γ_1	γ_2
	Scenario 1			
True values	1.000	0.000	1.000	0.000
True-ME	1.000 (0.007)	0.001 (0.007)	0.997 (0.085)	0.003 (0.088)
3Trees	1.000 (0.010)	0.001 (0.010)	0.997 (0.085)	0.003 (0.088)
	Scenario 2			
True values	1.000	0.000	1.000	0.000
True-ME	0.999 (0.011)	0.001 (0.007)	0.997 (0.085)	0.005 (0.137)
3Trees	0.999 (0.011)	0.001 (0.007)	0.998 (0.088)	−0.008 (0.144)
Linear-ME	2.396 (0.027)	0.001 (0.011)	0.996 (0.109)	−1.797 (0.114)
	Scenario 3			
True values	2.000	0.000	4.000	0.000
True-ME	2.000 (0.007)	0.001 (0.007)	3.997 (0.085)	0.003 (0.088)
3Trees	2.000 (0.017)	0.039 (0.578)	2.728 (0.119)	0.004 (0.113)
Linear-ME	2.001 (0.026)	0.001 (0.046)	2.727 (0.093)	0.003 (0.092)

the proposed algorithm with respect to the benchmark is confirmed in terms of average *clus*PMSE, with a slight increase in Monte Carlo variability.

Further settings are considered in Table 8 of Appendix A. Specifically, we report simulation results for five variations of Scenario 2. In Scenario 2A, we set the predictors' covariance at 0.75 instead of 0.4 to check the algorithm's behaviour in the case of highly correlated predictors. Scenario 2B considers the case of independent predictors instead. The 3Trees-Alg algorithm seems not influenced by the correlation between predictors, reporting similar performance in the two scenarios. Scenario 2C considers the case of a lower value of the intraclass correlation coefficient. In this case, the performance remains unchanged in terms of *clus*PMSE, but it is improved in terms of PMSE due to the reduced impact of the random component. Finally, in Scenarios 2D and 2E, the coefficients of the linear part increase to 5 or decrease to 0.5. Overall, the performance of the 3Trees algorithm is essentially unchanged. To check the behaviour of the procedures with more sample sizes, Table 9 in Appendix A presents the results of the same study as Table 2, where other eight non-significant explanatory variables were added to Scenario 2. The procedure is still able to capture which variables have no linear effect in this quasi-linear scenario.

4.2 Inferential accuracy

The Algorithm 3Trees-Alg-PSI (Algorithm 2 in Sect. 3.1) is devised to provide valid confidence intervals. We run a further simulation study applying the 3Trees-Alg-PSI

Table 3 Monte Carlo average length and Monte Carlo coverage of the 95% confidence intervals for the parameters of the linear component and for the random effects standard deviation. (Scenario 2 $J = 500$, $n_j = 100$, number of Monte Carlo runs: 500)

Algorithms	Parameters				
	σ_u	β_1	β_2	γ_1	γ_2
	Monte Carlo mean of the interval length				
True-ME	0.215	0.030	0.019	0.333	0.524
Split-True-ME	0.304	0.043	0.027	0.472	0.743
3Trees-Alg-PSI	0.305	0.043	0.027	0.521	0.790
	Monte Carlo coverage				
True-ME	0.934	0.946	0.964	0.948	0.948
Split-True-ME	0.960	0.960	0.976	0.954	0.958
3Trees-Alg-PSI	0.948	0.958	0.976	0.952	0.946

Algorithm to evaluate its performance, using 500 data sets of $J = 500$ clusters of size $n_j = 100$.

For comparison, we compute the confidence intervals for the parameters of interest using a correctly-specified mixed effect model (true model) fitted either on the entire data set (True-ME) and on the same split sample used in the Estimation step of Algorithm 3Trees-Alg-PSI, labelled Split-True-ME. We compare the Monte Carlo averages of the interval length and the Monte Carlo actual coverage. We are reporting here the results for Scenario 2 (Table 3), while those for Scenario 1 and 3 are in Appendix A (Table 10).

Note that the 3Trees-Alg-PSI is directly comparable with Split-True-ME, as they both are fitted on half of the data set. In summary, the length of the confidence intervals obtained with the 3Trees-Alg-PSI is in line with a mixed effect model fitted on the halved sample (Split-True-ME). Instead, the confidence intervals obtained by True-ME are shorter since they are obtained using the whole data set. Therefore, the loss of efficiency is attributable to the reduced sample size and not to the tree selection procedure. Moreover, 3Trees-Alg-PSI provides confidence intervals of the expected coverage, in line with correctly specified mixed effect models, as if the regression function was known. The results for Scenario 1 (Table 10) are in line with those of Scenario 2. As expected, for Scenario 3 the performances are worse. Deeper trees are required for this data generating process complexity, as highlighted by simulations in Table 1.

5 An illustrative example on invalsi tests in italian schools

We apply a 3Trees model to data from the Italian Invalsi test (see, e.g. Cardone et al. 2019). These data concern students who participated in the Invalsi Math tests while attending 5^{th} grade in 2013–2014 and then attending the 8^{th} grade in 2016–2017. The Invalsi data have a multilevel structure where the individual-level units are represented

Table 4 Student and school level variables

Pupil-level variables	
MATH8	*Response*: Test score at 8th grade (0-100)
MATH5	Test score at 5th grade (0-100)
SES	Socio-economic status (standardized)
GENDER	Gender (0 = male, 1 = female)
REGULAR	Regular career (0 = regular, 1 = one year early, 2 = one year late)
IMM	Citizenship (0 = Italian, immigrant: 1 = 1st generation, 2 = 2nd generation)
School-level variables	
AREA	Geographical area (0 = NE, 5 categories)
TOWN	School located in the provincial capital (1 = yes, 0 = no)
CLSIZE	Average number of students per class
SCSIZE	Number of classes in the school
SCTYPE	Type of school (0 = non-public, 1 = public)

by pupils and the cluster-level units by the schools, grouping the students. The data set contains 409 528 pupils in 5 777 Italian schools, with an average number of tested students per school of 70.94. Here the aim is to predict the Math achievement at 8^{th} grade, having some clues on understanding why results differ. The student and school-level predictors are described in Table 4.

We exploit the 3Trees model (4) with pupils as level 1 units and schools as level 2 units, where the random effect part is aimed to capture unobserved factors at the school-level. We first use the 3Trees-Alg (Algorithm 1) to predict the Math score at the 8^{th} grade. Table 5 compares the performance of the 3Trees model using Algorithm 1 (with tree depths set at 3) and the same algorithms considered in the simulation study. The 3Trees-Alg has the best performance for predicting the math score of a student in an observed school ($clus$PMSE) and predicting the response of a new student in a new school (PMSE). The RE-EM provides the second-best prediction when `cpmin`= 0.0001, providing a tree with 55 terminal nodes. The performance when `cpmin`= 0.001 is worse, but the resulting tree, with 15 terminal nodes, is more interpretable. The CART algorithm, ignoring the two-level structure of the data, performs worse than 3Trees and Linear-ME but avoids overfitting, thanks to the pruning procedure based on cross-validation.

The three methods accounting for clustering show a reduction of the PMSE when using the random effects to predict the response for a new unit in an observed cluster ($clus$PMSE). For the 3Trees model, this reduction is about 12%, which is much smaller than what was observed in the simulations (Table 1) due to the lower value of the ICC (0.13 versus 0.75).

In order to make inference on the coefficients of the predictors, we apply the 3Trees-Alg-PSI (Algorithm 2), which randomly splits the schools into two independent sub-samples for post-selection inference. Estimating the contextual effects of socio-economic status (SES) and mathematical background (MATH5) of pupils belonging to the same school is relevant for inferential purposes. To this end, we add

Table 5 Comparison of the the predictive performance for Math score in Invalsi data (*clus*MSE and MSE on train data, *clus*PMSE and PMSE on test data)

Algorithm	*clus*MSE	MSE	*clus*PMSE	PMSE
3Trees	214.13	251.77	220.15	250.13
RE-EM (`cpmin`= 0.001)	219.32	259.23	225.97	258.09
RE-EM (`cpmin`= 0.0001)	214.30	252.64	220.67	251.35
Linear RE	220.46	259.34	226.86	257.83
CART		270.84		270.49

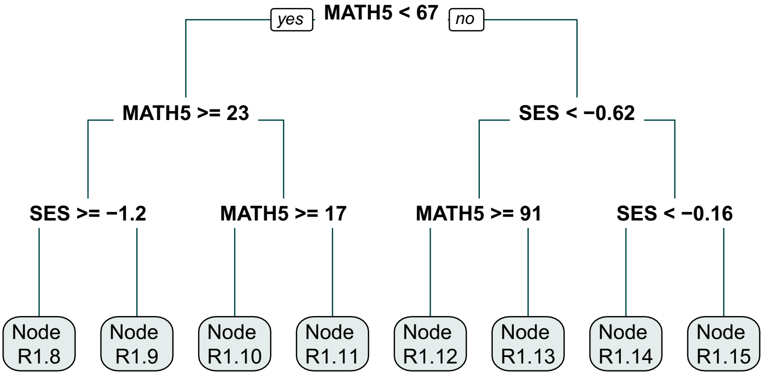


Fig. 3 First tree of the 3Trees model on Invalsi data using Algorithm 2

to the predictors the school averages of these two variables, named gr.M(SES) and gr.M(MATH5). To facilitate the interpretation of the results and keep control of the maximum size of the selected model, we set the maximum tree depth at three.

We first describe the tree component in the 3Trees model selected by 3Trees-Alg-PSI. The algorithm represents the predictors' partition of each tree as a factor with M_t labels corresponding to the partition regions. The factor is then parameterized as $(M_t - 1)$ dummy variables for the usual identifiability constraints. The first region plays as the reference. For simplicity, in the factor description, T stands for Tree and R for Region. The regions selected on the student-level predictors by the first tree are depicted in Fig. 3 and summarized as follows.

$$\begin{array}{ll}
\text{T1 R1.8 (reference):} & \mathbb{I}\,(23 \leq \text{MATH5} < 67\ \&\ \text{SES} \geq -1.21) \\
\text{T1 R1.9 :} & \mathbb{I}\,(23 \leq \text{MATH5} < 67\ \&\ \text{SES} < -1.21) \\
\text{T1 R1.10 :} & \mathbb{I}\,(17 \leq \text{MATH5} < 23) \\
\text{T1 R1.11 :} & \mathbb{I}\,(\text{MATH5} < 17) \\
\text{T1 R1.12 :} & \mathbb{I}\,(\text{MATH5} \geq 91\ \&\ \text{SES} < -0.62) \\
\text{T1 R1.13 :} & \mathbb{I}\,(23 \leq \text{MATH5} < 67\ \&\ \text{SES} < -0.62) \\
\text{T1 R1.14 :} & \mathbb{I}\,(\text{MATH5} \geq 67\ \&\ -0.62 \leq \text{SES} < -0.16) \\
\text{T1 R1.15 :} & \mathbb{I}\,(\text{MATH5} \geq 67\ \&\ \text{SES} \geq -0.16)
\end{array}$$

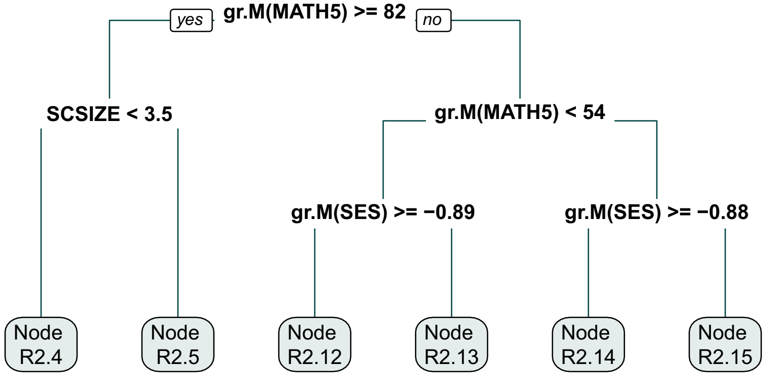


Fig. 4 Second tree of the 3Trees model on Invalsi data using Algorithm 2

The second tree, on the school-level predictors, is depicted in Fig. 4 and its selected regions are listed below.

$$\begin{array}{ll}
\text{T2 R2.4 (reference)}: & \mathbb{I}\,(\text{gr.M(MATH5)} \geq 82 \ \& \ \text{SCSIZE} < 3.5) \\
\text{T2 R2.5}: & \mathbb{I}\,(\text{gr.M(MATH5)} \geq 82 \ \& \ \text{SCSIZE} \geq 3.5) \\
\text{T2 R2.12}: & \mathbb{I}\,(\text{gr.M(MATH5)} < 54 \ \& \ \text{gr.M(SES)} \geq -0.89) \\
\text{T2 R2.13}: & \mathbb{I}\,(\text{gr.M(MATH5)} < 54 \ \& \ \text{gr.M(SES)} < -0.89) \\
\text{T2 R2.14}: & \mathbb{I}\,(54 \leq \text{gr.M(MATH5)} < 82 \ \& \ \text{gr.M(SES)} \geq -0.88) \\
\text{T2 R2.15}: & \mathbb{I}\,(54 \leq \text{gr.M(MATH5)} < 82 \ \& \ \text{gr.M(SES)} < -0.88)
\end{array}$$

Finally, the third tree jointly considers all predictors. The selected regions are depicted in Fig. 5 and listed below.

$$\begin{array}{ll}
\text{T3 R2.8 (ref.)}: & \mathbb{I}\,(35 \leq \text{MATH5} < 77 \ \& \ \text{AREA} \neq \text{South, Islands}) \\
\text{T3 R3.9}: & \mathbb{I}\,(35 \leq \text{MATH5} < 77 \ \& \ \text{AREA} = \text{South, Islands}) \\
\text{T3 R3.10}: & \mathbb{I}\,(\text{MATH5} \geq 77 \ \& \ \text{AREA} = \text{South, Islands}) \\
\text{T3 R3.11}: & \mathbb{I}\,(\text{MATH5} < 77 \ \& \ \text{AREA} \neq \text{South, Islands}) \\
\text{T3 R3.12}: & \mathbb{I}\,(\text{MATH5} < 35 \ \& \ \text{AREA} = \text{Center}) \\
\text{T3 R3.13}: & \mathbb{I}\,(\text{MATH5} < 35 \ \& \ \text{AREA} = \text{Nord–West, Nord–East}) \\
\text{T3 R3.14}: & \mathbb{I}\,(\text{MATH5} < 35 \ \& \ \text{AREA} = \text{South, Islands} \ \& \ \text{gr.M(MATH5} < 64)) \\
\text{T3 R3.15}: & \mathbb{I}\,(\text{MATH5} < 35 \ \& \ \text{AREA} = \text{South, Islands} \ \& \ \text{gr.M(MATH5} \geq 64))
\end{array}$$

The parameter estimates and their confidence intervals are reported in Table 6 in the column labelled 3Trees model. As expected, the Math score at 5^{th} grade (MATH5) is a key predictor, with a significant main effect in the linear component, in addition to non-linear and interaction effects captured by the first and third trees. Concerning other student-level predictors, it is worth noting that one of the most relevant predictors is the socio-economic status, confirming an enduring issue and the necessity of additional support for children with low socio-economic status in order to overcome social inequalities.

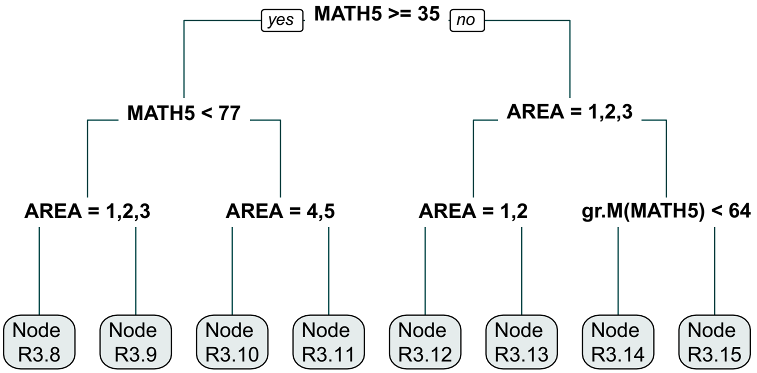


Fig. 5 Third tree of the 3Trees model on Invalsi data using Algorithm 2

Regarding the school-level predictors, the coefficient of type of school (SCTYPE) is negative. Thus, conditionally on the other predictors, students of public schools have a lower performance.

Looking at the confidence intervals, we note that two school-level predictors, namely the number of classes in the school (CLSIZE) and the school's location (TOWN), have confidence intervals crossing zero. In addition, the tree for the school-level predictors has two regions (R2.12 and R2.13) whose confidence intervals for the coefficients largely overlap. Similarly, in the third tree, the regions R3.10 and R3.11. Consequently, these two trees can be pruned. Table 6 also reports the parameter estimates obtained from a reduced model where the trees are pruned and the non-significant predictors SCSIZE and TOWN are omitted. This reduced model has four parameters less than the full model and is nested in it. The pruned 3Trees model is preferred based on the Likelihood Ratio test comparing the two models (statistic $= 4.67$, df $= 4$, p-value $= 0.322$).

6 Conclusions

Tree-based regression models are a class of predictive algorithms that are conceptually simple and able to take into account both interactions and non-linearities. However, they struggle with linear dependencies and assume *iid* data.

This paper proposes a multilevel extension of regression trunk models where the tree component consists of multiple trees. In particular, we call 3Trees model a mixed-effect model with a linear part and three trees. The trees are aimed to take into account non-linearities and interactions among individual-level predictors, cluster-level predictors and cross-levels. The model can be easily extended to a larger number of trees.

The key idea behind our proposal is to develop a flexible class of models capable of accurately approximating the true regression function while preserving interpretability. Indeed, the limitation of a standard regression tree is that, even if it is able to reconstruct a regression function in the presence of non-linear effects and interactions, it does not directly reveal which effect is linear and which is not, which predictor is involved in an

Table 6 Estimates and 95% confidence intervals for the 3Trees model using Algorithm 2 and the corresponding pruned tree

	3Trees model		Pruned 3Trees model	
MATH5	0.626	(0.616, 0.636)	0.626	(0.616, 0.636)
SES	2.129	(2.028, 2.230)	2.130	(2.029, 2.231)
GENDER (female)	− 1.434	(−1.564, −1.304)	−1.434	(−1.564, −1.304)
REGULAR (before)	1.648	(1.064, 2.232)	1.652	(1.068, 2.236)
REGULAR (after)	−3.554	(−4.116, −2.992)	−3.554	(−4.116, 2.991)
IMM (1st gen)	−1.215	(−1.703, −0.728)	−1.216	(−1.703, −0.728)
IMM (2nd gen)	−1.485	(−1.796, −1.175)	−1.485	(−1.795, −1.174)
AREA (North-West)	0.370	(−0.351, 1.092)	0.383	(−0.338, 1.105)
AREA (Centre)	−2.070	(−2.777, −1.364)	−2.037	(−2.738, −1.336)
AREA (South)	7.053	(5.907, 8.199)	7.084	(5.941, 8.227)
AREA (Islands)	8.365	(7.202, 9.528)	8.383	(7.223, 9.544)
TOWN	0.273	(−0.292, 0.837)		
CLSIZE	0.148	(0.123, 0.172)	0.149	(0.125, 0.173)
SCSIZE	0.038	(−0.075, 0.151)		
SCTYPE (public)	−1.702	(−2.804, −0.600)	−1.665	(−2.645, −0.686)
School Mean gr.M(SES)	1.180	(0.456, 1.904)	1.167	(0.533, 1.801)
School Mean gr.M(MATH5)	−0.302	(−0.345, −0.258)	−0.302	(−0.345, −0.258)
First tree, region R1.9	0.875	(0.524, 1.225)	0.883	(0.532, 1.234)
First tree, region R1.10	5.547	(3.966, 7.129)	5.554	(3.973, 7.136)
First tree, region R1.11	13.639	(10.669, 16.609)	13.641	(10.671, 16.611)
First tree, region R1.12	−3.031	(−3.689, −2.374)	−3.028	(−3.685, −2.371)
First tree, region R1.13	−0.529	(−0.865, −0.193)	−0.527	(−0.863, −0.191)
First tree, region R1.14	0.503	(0.166, 0.840)	0.503	(0.166, 0.840)
First tree, region R1.15	1.560	(1.298, 1.822)	1.561	(1.299, 1.823)
Second tree, region R2.5	3.086	(0.178, 5.995)	3.170	(0.270, 6.069)
Second tree, region R2.12	−2.334	(−5.711, 1.043)		
Second tree, region R2.13	0.698	(−3.499, 4.896)		
Second tree, region R2.14	0.176	(−2.263, 2.614)		
Second tree, region R2.15	0.292	(−2.741, 3.326)		
Second tree, region R2.12 + R2.13			−1.299	(−4.505, 1.907)
Second tree, region R2.14 + R2.15			0.312	(−2.110, 2.734)
Third tree, region R3.9	−7.246	(−8.136, −6.356)	−7.246	(−8.136, −6.356)
Third tree, region R3.10	−12.352	(−13.312,−11.391)	−12.351	(−13.312,−11.391)
Third tree, region R3.11	3.977	(3.727, 3.727)	3.978	(3.728, 4.228)
Third tree, region R3.12	−1.102	(−1.865, −0.339)	−1.103	(−1.866, −0.340)

Table 6 continued

	3Trees model		Pruned 3Trees model	
Third tree, region R3.13	2.019	(0.960, 3.077)	2.015	(0.956, 3.073)
Third tree, region R3.14	−2.163	(−3.288, −1.038)	−2.155	(−3.279, −1.030)
Constant	32.595	(28.085, 37.106)	32.643	(28.159, 37.128)
Level 2 (school) std deviation	6.004	(5.830, 6.186)	6.011	(5.836, 6.192)
Level 1 (pupil) std deviation	14.752	(14.706, 14.797)	14.752	(14.706, 14.797)

interaction and which is not. As explained by Gottard et al. (2020), in most tree-based algorithms, including CART and random forests, the ordering of the splitting variables and their importance can be different from the ordering and the importance of the direct effects on the response. The interpretation of a tree is in terms of predictive power and not on *attribution* (Efron 2020) and direct effects as in traditional regression. From this point of view, 3Trees models constitute an interesting bridge between a proper statistical model and a machine learning algorithm.

To estimate the 3Trees model parameters, we introduce two iterative algorithms. The first algorithm is specific for predictive purposes and showed a good predictive performance in the simulation study and the applied study on Invalsi data. It produces transparent and interpretable predictions. However, this algorithm does not provide valid confidence intervals, as the shape of the regression function is learned during the estimation process. We propose a post-selection inference procedure that provides valid confidence intervals for the coefficients and predicted values to overcome this limitation. This procedure is based on the splitting sample procedure, which is easy to apply and shows good performance if the number of clusters in the data is not too small. Notice that, in likelihood-based inference for multilevel models, the number of clusters is crucial for estimating the level 2 variance σ_u^2, which in turn affects the standard errors of regression coefficients, especially those related to level 2 predictors (Elff et al. 2021). In the case of a few clusters, say less than 30 in each split sample, the level 2 variance is appreciably underestimated. Thus the confidence intervals are shorter than due. This issue can be addressed by REML estimation, though this comes at the cost of losing efficiency and it prevents using LR tests for pruning.

The proposed algorithms for 3Trees models can be easily extended to allow for predictors with random slopes (see Appendix B). Indeed, there is no need to modify the algorithm but to specify the random component properly. It is worth noting that the role of a random slope is to account for a large unexplained between-cluster variation in the regression coefficient of a level 1 predictor. This scenario is less likely in a 3Trees model due to the flexibility of the tree component, notably for what concerns cross-level interactions. It can happen that the increase in complexity induced by random slopes may not be adequately rewarded in terms of predictive power. It will be up to the researcher to decide whether to adopt a random slope or leave cross-level interactions to be explained by the trees. In general, random slopes should be considered only for predictors playing a key role in the research aims.

Finally, further work is needed to apply the proposed 3Trees model and related algorithms to high dimensional settings. When the number of predictors increases, the proposed algorithms can be modified to deal with the high dimensionality issue. For instance, one could replace the likelihood-based estimation procedure with the lasso-type estimator of Groll and Tutz (2014). In addition, recently Rügamer et al. (2022) proposed a procedure for post-selection inference for linear mixed-effect models and additive models, implementing a multi-stage selection procedure. Further research is needed to evaluate if their proposal can be adapted to the case of 3Trees models.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: Further results of the Monte Carlo study

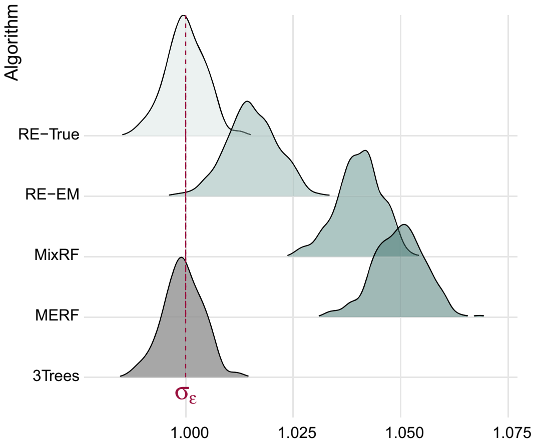


Fig. 6 Monte Carlo distribution of the estimate of the parameter for the residual standard deviation σ_ε for the 3Trees and RE-EM algorithms, compared with the mixed-effect model with the *true* (linear) regression functions. Scenario 1 of Table 1

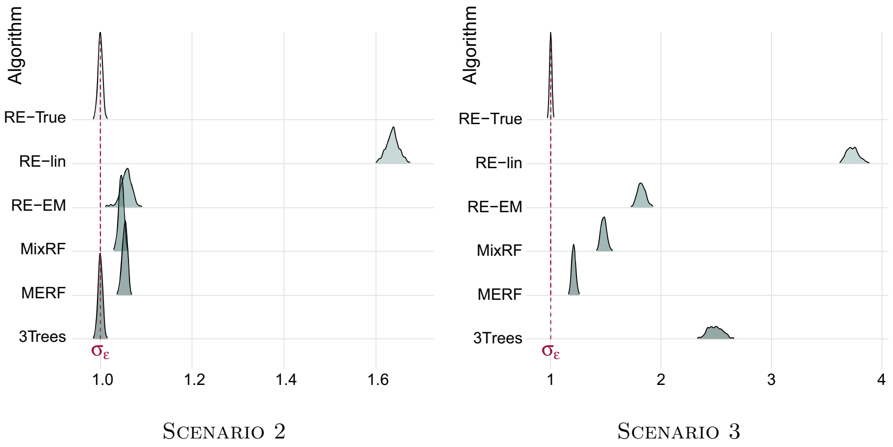


Fig. 7 Monte Carlo distributions of the estimate the residual standard deviation σ_ε for the 3Trees and RE-EM algorithms, compared with the mixed-effect models with the *true* and the linear regression functions respectively. Scenarios 2 and 3 of Table 1

Table 7 Monte Carlo averages and standard deviations of the Mean Squared Error and Predictive Mean Squared Error, computed with fixed effects only (MSE, PMSE), or including BLUP predictions of the random effects ($clus$MSE and $clus$PMSE). Scenarios 1, 2 and 3 ($J = 100$, $n_j = 15$ for both the training set and the test set, number of Monte Carlo runs: 500)

Algorithm	$clus$MSE	MSE	$clus$PMSE	PMSE
	Scenario 1			
True-ME	0.932 (0.035)	3.907 (0.475)	1.067 (0.040)	3.912 (0.473)
3Trees	0.917 (0.035)	3.846 (0.499)	1.084 (0.041)	3.883 (0.497)
RE-EM	0.962 (0.045)	3.653 (0.533)	2.890 (0.151)	5.604 (0.586)
MixRF	0.316 (0.014)	3.231 (0.475)	1.180 (0.044)	4.071 (0.472)
MERF	0.247 (0.020)	2.384 (0.661)	1.191 (0.049)	3.418 (0.570)
CART		3.275 (0.372)		3.580 (0.396)
STIMA		3.565 (0.437)		3.593 (0.437)
	Scenario 2			
True-ME	0.931 (0.035)	3.874 (0.472)	1.068 (0.040)	3.882 (0.470)
3Trees	1.050 (0.185)	4.296 (0.913)	1.227 (0.219)	4.328 (0.926)
RE-EM	1.268 (0.065)	4.946 (0.822)	1.476 (0.081)	4.986 (0.817)
MixRF	0.347 (0.014)	4.299 (0.563)	1.267 (0.051)	5.256 (0.574)
MERF	0.253 (0.016)	2.630 (0.676)	1.214 (0.049)	3.669 (0.612)
Linear-ME	2.501 (0.116)	7.347 (0.723)	2.858 (0.134)	7.369 (0.728)
CART		4.124 (0.424)		4.259 (0.442)
STIMA		4.113 (0.466)		4.159 (0.476)

Table 7 continued

Algorithm	*clus*MSE	MSE	*clus*PMSE	PMSE
	SCENARIO 3			
True-ME	0.931 (0.035)	3.906 (0.475)	1.069 (0.039)	3.913 (0.473)
3Trees	6.360 (0.921)	9.360 (1.062)	7.659 (0.964)	9.958 (1.045)
RE-EM	6.260 (0.819)	11.503 (1.383)	8.239 (0.951)	12.854 (1.393)
MixRF	1.326 (0.185)	6.386 (0.947)	4.450 (0.560)	9.571 (1.172)
MERF	0.595 (0.105)	4.454 (1.588)	2.909 (0.393)	6.860 (1.569)
Linear-ME	13.176 (1.278)	16.784 (1.427)	14.673 (1.206)	16.900 (1.302)
CART		9.339 (0.869)		11.243 (1.060)
STIMA		11.255 (1.158)		11.761 (1.071)

Table 8 Monte Carlo averages and standard deviations of the Mean Squared Error and Predictive Mean Squared Error, computed with fixed effects only (MSE, PMSE), or including BLUP predictions of the random effects (*clus*MSE and *clus*PMSE). Five variants of Scenario 2 ($J = 100$, $n_j = 15$ for both the training set and the test set, number of Monte Carlo runs: 500)

Algorithm	*clus*MSE	MSE	*clus*PMSE	PMSE
	SCENARIO 2A: HIGHLY CORRELATED PREDICTORS ($\rho = 0.75$)			
True-ME	0.931 (0.035)	3.873 (0.473)	1.069 (0.040)	3.881 (0.472)
3Trees	1.045 (0.181)	4.220 (0.820)	1.221 (0.215)	4.260 (0.822)
RE-EM	1.254 (0.072)	4.761 (0.670)	1.463 (0.088)	4.808 (0.668)
MixRF	0.336 (0.014)	4.008 (0.562)	1.242 (0.046)	4.925 (0.567)
MERF	0.246 (0.015)	2.811 (0.608)	1.208 (0.048)	3.822 (0.558)
Linear-ME	2.502 (0.119)	7.356 (0.728)	2.856 (0.134)	7.365 (0.740)
CART		4.078 (0.442)		4.222 (0.447)
STIMA		4.119 (0.470)		4.161 (0.475)
	SCENARIO 2B: INDEPENDENT PREDICTORS			
True-ME	0.931 (0.035)	3.874 (0.472)	1.069 (0.040)	3.882 (0.471)
3Trees	1.055 (0.189)	4.281 (0.970)	1.235 (0.225)	4.313 (0.968)
RE-EM	1.274 (0.067)	5.141 (0.958)	1.489 (0.085)	5.193 (0.948)
MixRF	0.359 (0.016)	4.146 (0.602)	1.265 (0.050)	5.097 (0.607)
MERF	0.267 (0.021)	2.214 (0.749)	1.229 (0.051)	3.298 (0.652)
Linear-ME	2.502 (0.116)	7.353 (0.743)	2.861 (0.138)	7.378 (0.736)
CART		4.176 (0.453)		4.318 (0.460)
STIMA		4.116 (0.468)		4.157 (0.475)
	SCENARIO 2C: INTRA- CLASS CORRELATION = 0.20			
True-ME	0.944 (0.036)	1.234 (0.061)	1.056 (0.038)	1.245 (0.058)
3Trees	1.081 (0.194)	1.469 (0.460)	1.230 (0.222)	1.503 (0.466)
RE-EM	1.286 (0.050)	2.282 (0.245)	1.484 (0.071)	2.327 (0.244)
MixRF	0.349 (0.015)	1.147 (0.186)	1.265 (0.051)	2.067 (0.200)

Table 8 continued

Algorithm	*clus*MSE	MSE	*clus*PMSE	PMSE
MERF	0.255 (0.012)	0.436 (0.055)	1.201 (0.047)	1.363 (0.075)
Linear-ME	2.509 (0.117)	4.681 (0.277)	2.851 (0.133)	4.708 (0.277)
CART		1.909 (0.120)		2.008 (0.139)
STIMA		1.497 (0.069)		1.519 (0.070)
	Scenario 2D: strong linear component ($\beta_1 = \gamma_1 = 5$)			
True-ME	0.931 (0.035)	3.874 (0.472)	1.068 (0.040)	3.882 (0.470)
3Trees	1.050 (0.185)	4.296 (0.913)	1.227 (0.219)	4.328 (0.926)
RE-EM	3.358 (0.828)	22.844 (6.030)	4.147 (0.981)	23.346 (5.977)
MixRF	0.854 (0.042)	15.020 (2.152)	3.004 (0.208)	17.590 (2.251)
MERF	0.323 (0.033)	9.145 (4.413)	1.547 (0.102)	10.770 (4.240)
Linear-ME	2.501 (0.116)	7.347 (0.723)	2.858 (0.134)	7.369 (0.728)
CART		12.185 (1.100)		13.595 (1.259)
STIMA		4.113 (0.466)		4.159 (0.476)
	Scenario 2E: weak linear component ($\beta_1 = \gamma_1 = 0.5$)			
True-ME	0.931 (0.035)	3.874 (0.472)	1.068 (0.040)	3.882 (0.470)
3Trees	1.050 (0.185)	4.296 (0.913)	1.227 (0.219)	4.328 (0.926)
RE-EM	1.024 (0.040)	4.244 (0.614)	1.191 (0.055)	4.272 (0.614)
MixRF	0.334 (0.014)	3.915 (0.505)	1.213 (0.048)	4.813 (0.517)
MERF	0.258 (0.020)	2.258 (0.734)	1.204 (0.050)	3.313 (0.635)
Linear-ME	2.501 (0.116)	7.347 (0.723)	2.858 (0.134)	7.369 (0.728)
CART		3.892 (0.437)		3.983 (0.447)
STIMA		4.113 (0.466)		4.159 (0.476)

Table 9 Monte Carlo averages and standard deviations (in parenthesis) for the parameters of the linear component in Scenario 2 with 8 noise predictors, J clusters and n_j observations (same for both training and test sets), 500 Monte Carlo runs

Parameter	True	$J = 500$ and $n_j = 50$		$J = 100$ and $n_j = 15$	
β_1	1.0	1.054	(0.038)	1.257	(0.038)
β_2	0.0	0.000	(0.008)	0.003	(0.008)
β_3	0.0	0.000	(0.008)	−0.002	(0.008)
β_4	0.0	0.000	(0.008)	−0.002	(0.008)
β_5	0.0	0.001	(0.008)	−0.001	(0.008)
β_6	0.0	0.000	(0.008)	−0.001	(0.008)
β_7	0.0	0.001	(0.008)	0.002	(0.008)
β_8	0.0	−0.001	(0.008)	0.002	(0.008)
β_9	0.0	0.000	(0.008)	−0.001	(0.008)
β_{10}	0.0	0.000	(0.008)	0.000	(0.008)
γ_1	1.0	0.993	(0.107)	0.979	(0.107)
γ_2	0.0	−0.321	(0.643)	−0.659	(0.643)

Table 10 Monte Carlo average length and Monte Carlo coverage of the 95% confidence intervals for the parameters of the linear component and for the random effects standard deviation. Data generating model of Scenario 1 and 3 ($J = 500$, $n_j = 100$, number of Monte Carlo runs: 500)

Algorithms	Parameters				
	σ_u	β_1	β_2	γ_1	γ_2
	Monte Carlo mean of the interval length				
	Scenario 1				
True-ME	0.215	0.019	0.019	0.333	0.333
Split-True-ME	0.304	0.027	0.027	0.472	0.472
3Trees-Alg-PSI	0.304	0.034	0.033	0.472	0.472
	Scenario 3				
True-ME	0.215	0.019	0.019	0.333	0.333
Split-True-ME	0.304	0.027	0.027	0.472	0.472
3Trees-Alg-PSI	0.310	0.072	0.126	0.546	0.542
	Monte Carlo coverage				
	Scenario 1				
True-ME	0.938	0.962	0.964	0.946	0.948
Split-True-ME	0.964	0.962	0.976	0.956	0.942
3Trees-Alg-PSI	0.964	0.970	0.976	0.956	0.942
	Scenario 3				
True-ME	0.940	0.960	0.962	0.946	0.948
Split-True-ME	0.962	0.962	0.974	0.956	0.942
3Trees-Alg-PSI	0.962	0.938	0.086	0.000	0.962

Appendix B: A random slope 3Trees model for INVALSI data

We show the results of a 3Trees model with a random intercept and a random slope on SES. The model is fitted on Invalsi data using Algorithm 2, 3Trees-Alg-PSI. The

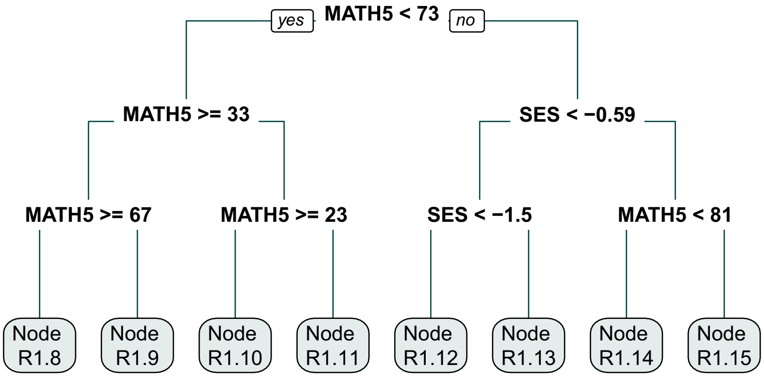


Fig. 8 First tree of the 3Trees model on Invalsi data using Algorithm 2 with random slope on SES

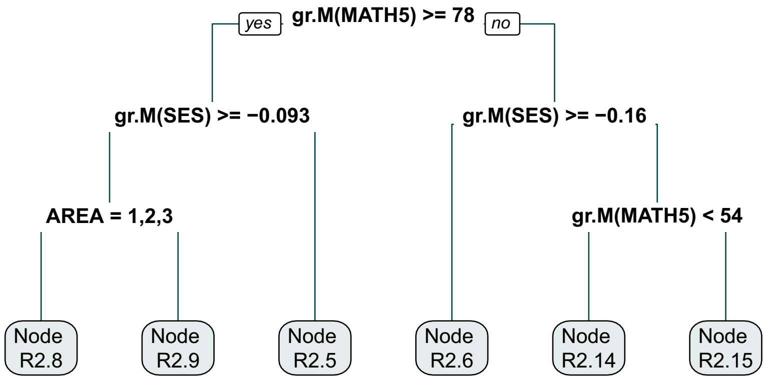


Fig. 9 Second tree of the 3Trees model on Invalsi data using Algorithm 2 with random slope on SES

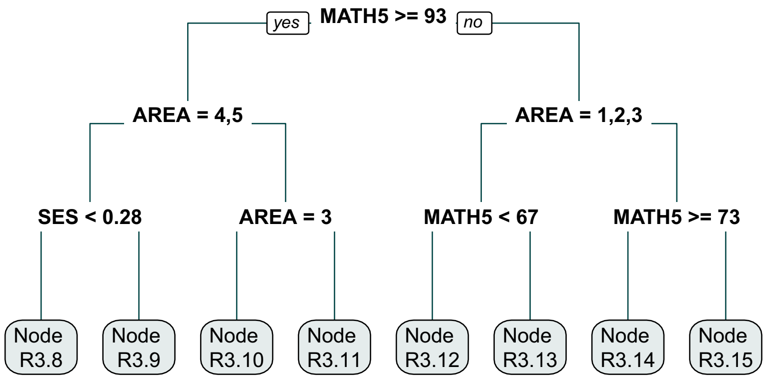


Fig. 10 Third tree of the 3Trees model on Invalsi data using Algorithm 2 with random slope on SES

algorithm is similar to the one for random intercept, but, both in the selection and in the estimation step, the specification of the random component in the mixed-effect model has been modified to include a random slope. In the following, we first describe the selected trees and then report the estimates table.

The regions selected on the student-level predictors by the first tree are depicted in Fig. 8 and listed below. Comparing this tree with the corresponding tree in the random intercept model (Fig. 3), we note that it involves the same predictors, namely MATH5 and SES. However, SES plays a less important role in the tree since it acts only when MATH $\geq$ 73. This is an expected consequence given that in the current model the effect of SES is modelled in a finer way by the random slope.

T1 R1.8 (ref.):	$\mathbb{I}\,(67 \leq \text{MATH5} < 73)$
T1 R1.9 :	$\mathbb{I}\,(33 \leq \text{MATH5} < 67)$
T1 R1.10 :	$\mathbb{I}\,(23 \leq \text{MATH5} < 33)$
T1 R1.11 :	$\mathbb{I}\,(\text{MATH5} < 23)$
T1 R1.12 :	$\mathbb{I}\,(\text{MATH5} \geq 73 \;\&\; \text{SES} < -1.50)$
T1 R1.13 :	$\mathbb{I}\,(\text{MATH5} \geq 73 \;\&\; -1.50 \leq \text{SES} < -0.59)$
T1 R1.14 :	$\mathbb{I}\,(73 \leq \text{MATH5} < 81 \;\&\; \text{SES} \geq -0.59)$
T1 R1.15 :	$\mathbb{I}\,(\text{MATH5} \geq 81 \;\&\; \text{SES} \geq -0.59)$

The second tree, dealing with the school-level predictors, is depicted in Fig. 9 and its selected regions are summarized as follows.

T2 R2.5 (ref.) :	$\mathbb{I}\,(\text{gr.M(MATH5)} \geq 78 \;\&\; \text{gr.M(SES)} < -0.093)$
T2 R2.6:	$\mathbb{I}\,(\text{gr.M(MATH5)} < 78 \;\&\; \text{gr.M(SES)} \geq -0.163)$
T2 R2.8 :	$\mathbb{I}\,(\text{gr.M(MATH5)} \geq 78 \;\&\; \text{gr.M(SES)} \geq -0.093 \;\&$ $\text{AREA} \neq \text{South, Islands})$
T2 R2.9 :	$\mathbb{I}\,(\text{gr.M(MATH5)} \geq 78 \;\&\; \text{gr.M(SES)} \geq -0.093 \;\&$ $\text{AREA} = \text{South, Islands})$
T2 R2.14 :	$\mathbb{I}\,(\text{gr.M(MATH5)} < 54 \;\&\; \text{gr.M(SES)} < -0.163)$
T2 R2.15 :	$\mathbb{I}\,(54 \leq \text{gr(MATH5)} < 78 \;\&\; \text{gr.M(SES)} < -0.163)$

Finally, the third tree jointly considers all predictors. The selected regions are depicted in Fig. 10 and described as follows.

T3 R2.8 (ref.):	$\mathbb{I}\,(\text{MATH5} \geq 93 \;\&\; \text{AREA} = \text{South, Islands} \;\&\; \text{SES} < 0.28)$
T3 R3.9 :	$\mathbb{I}\,(\text{MATH5} \geq 93 \;\&\; \text{AREA} = \text{South, Islands} \;\&\; \text{SES} \geq 0.28)$
T3 R3.10:	$\mathbb{I}\,(\text{MATH5} \geq 93 \;\&\; \text{AREA} = \text{Center})$
T3 R3.11 :	$\mathbb{I}\,(\text{MATH5} \geq 93 \;\&\; \text{AREA} = \text{Nord–West})$
T3 R3.12 :	$\mathbb{I}\,(\text{MATH5} < 67 \;\&\; \text{AREA} \neq \text{South, Islands})$
T3 R3.13 :	$\mathbb{I}\,(67 \leq \text{MATH5} < 93 \;\&\; \text{AREA} \neq \text{South, Islands})$
T3 R3.14 :	$\mathbb{I}\,(73 \leq \text{MATH5} < 93 \;\&\; \text{AREA} = \text{South, Islands})$
T3 R3.15 :	$\mathbb{I}\,(\text{MATH5} < 73 \;\&\; \text{AREA} = \text{South, Islands})$

Estimates and confidence intervals are reported in the first column of Table 11 (3Trees model). The estimates for the pruned version are presented in the second column (Pruned 3Trees model). According to the LR test (statistic $= 4.543$, df $= 3$, p-value $= 0.208$) the pruned version is to be preferred.

Table 11 Estimates and 95% confidence intervals for a 3Trees model with random slope on SES using Algorithm 2 and the corresponding pruned tree

	3Trees model		Pruned 3Trees model	
MATH5	0.609	(0.598, 0.619)	0.609	(0.598, 0.619)
SES	2.118	(2.022, 2.213)	2.117	(2.022, 2.212)
GENDER (female)	−1.503	(−1.632, −1.375)	−1.504	(−1.632, −1.375)
REGULAR (before)	1.780	(1.200, 2.360)	1.780	(1.200, 2.360)
REGULAR (after)	−3.931	(−4.489, −3.372)	−3.931	(−4.489, −3.373)
IMM (1st gen)	−1.393	(−1.884, −0.902)	−1.393	(−1.884, −0.902)
IMM (2nd gen)	−1.641	(−1.949, −1.333)	−1.640	(−1.948, −1.332)
AREA (North−West)	0.258	(−0.435, 0.951)	0.261	(−0.432, 0.954)
AREA (Centre)	−1.469	(−2.160, −0.778)	−1.470	(−2.157, −0.783)
AREA (South)	−12.410	(−13.329,−11.491)	−12.380	(−13.299,−11.463)
AREA (Islands)	−12.083	(−13.021,−11.495)	−12.065	(−13.000,−11.130)
TOWN	0.046	(−0.509, 0.600)		
CLSIZE	0.154	(0.129, 0.178)	0.153	(0.129, 0.178)
SCSIZE	0.123	(0.016, 0.231)	0.120	(0.013, 0.226)
SCTYPE (public)	−2.450	(−3.498, −1.403)	−2.349	(−3.387, −1.312)
School Mean gr.M(SES)	1.093	(0.259, 1.926)	1.288	(0.500, 2.075)
School Mean gr.M(MATH5)	−0.210	(−0.257, −0.164)	−0.206	(−0.250, −0.162)
First tree, region R1.9	2.187	(1.797, 2.577)	2.185	(1.796, 2.575)
First tree, region R1.10	5.774	(5.037, 6.510)	5.769	(5.033, 6.505)
First tree, region R1.11	13.480	(11.971, 14.989)	13.473	(11.964, 14.982)
First tree, region R1.12	0.309	(−0.313, 0.931)	0.303	(−0.319, 0.925)
First tree, region R1.13	1.436	(1.046, 1.827)	1.433	(1.043, 1.824)
First tree, region R1.14	2.431	(2.131, 2.730)	2.430	(2.131, 2.730)
First tree, region R1.15	3.747	(3.418, 4.075)	3.746	(3.417, 4.074)
Second tree, region R2.5	0.325	(−2.147, 2.798)		
Second tree, region R2.6	3.139	(1.373, 4.906)	3.213	(1.456, 4.970)
Second tree, region R2.9	2.717	(0.361, 5.073)		
Second tree, region R2.14	1.772	(−1.247, 4.790)		
Second tree, region R2.15	2.699	(0.765, 4.634)		
Second tree, region R2.5 + R2.9			1.642	(−0.457, 3.742)
Second tree, region R2.14 + R2.15			2.855	(0.929, 4.782)
Third tree, region R3.9	2.800	(2.024, 3.575)	2.812	(2.036, 3.587)
Third tree, region R3.10	−1.519	(−2.193, −0.845)	−1.524	(−2.199, −0.850)
Third tree, region R3.11	3.275	(2.784, 3.766)	3.274	(2.783, 3.765)
Third tree, region R3.12	−4.825	(−5.257, −4.394)	−4.823	(−5.254, −4.391)
Third tree, region R3.14	5.282	(4.714, 5.850)	5.289	(4.721, 5.857)
Third tree, region R3.15	10.011	(9.329, 10.694)	10.015	(9.333, 10.698)
Constant	26.834	(22.628, 31.040)	26.332	(22.311, 37.128)

Table 11 continued

	3Trees model		Pruned 3Trees model	
Level 2: intercept std deviation	5.855	(5.683 , 6.034)	5.861	(5.689 , 6.040)
Level 2: slope std deviation	1.103	(0.993 , 1.212)	1.103	(0.993 , 1.212)
Level 2: intercept-slope correlation	−0.200	(−0.281 , −0.113)	−0.200	(−0.279 , −0.111)
Level 1: intercept std deviation	14.597	(14.552 , 14.642)	14.597	(14.552 , 14.642)

References

Abdolell M, LeBlanc M, Stephens D, Harrison R (2002) Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Stat Med 21(22):3395–3409

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48

Bauer DJ, Curran PJ (2005) Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. Multivar Behav Res 40(3):373–400

Benjamini Y (2010) Simultaneous and selective inference: Current successes and future challenges. Biom J 52(6):708–721

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. Ann Stat 41(2):802–837

Breiman L, Friedman J, Stone C, Olshen R (1984) Classification and regression trees. CRC Press, Boca Raton, FL

Bryk AS, Raudenbush SW (2001) Hierarchical linear models: Applications and data analysis methods, 2nd edn. Sage Publications Inc, Thousand Oaks, CA

Buja A, Brown L, Berk R, George E, Pitkin E, Traskin M, Zhang K, Zhao L (2019) Models as approximations i: Consequences illustrated with linear regression. Stat Sci 34(4):523–544

Capitaine L, Genuer R, Thiébaut R (2021) Random forests for high-dimensional longitudinal data. Stat Methods Med Res 30(1):166–184

Cardone M, Falzetti P, Sacco C (2019) Invalsi data for school system improvement: the value added. Working Papers INVALSI, 43/2019 [Online]

Cox DR (1975) A note on data-splitting for the evaluation of significance levels. Biometrika 62(2):441–444

Dusseldorp E, Conversano C, Van Os BJ (2010) Combining an additive and tree-based regression model simultaneously: Stima. J Comput Graph Stat 19(3):514–530

Dusseldorp E, Meulman JJ (2004) The regression trunk approach to discover treatment covariate interaction. Psychometrika 69(3):355–374

Efron B (2020) Prediction, estimation, and attribution. Int Stat Rev 88:S28–S59

Elff M, Heisig JP, Schaeffer M, Shikano S (2021) Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. British Journal of Political Science 51(1):412–426

Eo S-H, Cho H (2014) Tree-structured mixed-effects regression modeling for longitudinal data. J Comput Graph Stat 23(3):740–760

Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H (2018) Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behav Res Methods 50(5):2016–2034

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics. Springer, New York

Fu W, Simonoff JS (2015) Unbiased regression trees for longitudinal and clustered data. Computational Statistics & Data Analysis 88:53–74

Gottard A, Vannucci G, Marchetti GM (2020) A note on the interpretation of tree-based regression models. Biom J 62(6):1564–1573

Groll A, Tutz G (2014) Variable selection for generalized linear mixed models by l 1-penalized estimation. Stat Comput 24(2):137–154

Hajjem A, Bellavance F, Larocque D (2011) Mixed effects regression trees for clustered data. Statistics & Probability Letters 81:451–459

Hajjem A, Bellavance F, Larocque D (2014) Mixed-effects random forest for clustered data. J Stat Comput Simul 84:1–18

Hajjem A, Larocque D, Bellavance F (2017) Generalized mixed effects regression trees. Statistics & Probability Letters 126:114–118

Hiabu M, Nielsen JP, Scheike TH (2021) Nonsmooth backfitting for the excess risk additive regression model with two survival time scales. Biometrika 108(2):491–506

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 15(3):651–674

Loh W-Y (2002) Regression tress with unbiased variable selection and interaction detection. Stat Sin 12:361–386

Loh W-Y, Zheng W (2013) Regression trees for longitudinal and multiresponse data. The Annals of Applied Statistics 7:495–522

Miller P, McArtor D, Lubke G (2017) metboost: Exploratory regression analysis with hierarchically clustered data. arXiv:1702.03994v1 [stat.ML]

Pellagatti M, Masci C, Ieva F, Paganoni AM (2021) Generalized mixed-effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining: The ASA Data Science Journal 14(3):241–257

Pinheiro J, Bates D (2006) Mixed-effects models in S and S-PLUS. Springer Science & Business Media, Berlin

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rinaldo A, Wasserman L, G'Sell M (2019) Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. Ann Stat 47(6):3438–3469

Robinson GK (1991) That BLUP is a Good Thing: The Estimation of Random Effects. Stat Sci 6(1):15–32

Rügamer D, Baumann PF, Greven S (2022) Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis 167*, in press

Segal MR (1992) Tree-structured methods for longitudinal data. J Am Stat Assoc 87(418):407–418

Seibold H, Hothorn T, Zeileis A (2019) Generalised linear model trees with global additive effects. Adv Data Anal Classif 13(3):703–725

Sela R, Simonoff J (2012) Re-em trees: A data mining approach for longitudinal and clustered data. Mach Learn 86(2):169–207

Sela RJ, Simonoff JS(2021) REEMtree: Regression Trees with Random Effects. R package version 0.90.4

Skrondal A, Rabe-Hesketh S (2009) Prediction in multilevel generalized linear models. J R Stat Soc A Stat Soc 172(3):659–687

Snijders T, Bosker R (2012) Multilevel analysis: An introduction to basic and advanced multilevel modeling, 2nd edn. SAGE Publications Inc, London

Therneau T, Atkinson B (2019) rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15

Wang J, Gamazon ER, Pierce BL, Stranger BE, Im HK, Gibbons RD, Cox NJ, Nicolae DL, Chen LS (2016) Imputing gene expression in uncollected tissues within and beyond gtex. The American Journal of Human Genetics 98(4):697–708

Wermuth N, Cox D (1998) On association models defined over independence graphs. Bernoulli 4(4):477–495

Zhang H (1998) Classification trees for multiple binary responses. J Am Stat Assoc 93(441):180–193