# Experimental Assessment of Heterogeneous Fuzzy Regression Trees

José Luis Corcuera Bárcena[a], Pietro Ducange[b], Riccardo Gallo, Francesco Marcelloni[c], Alessandro Renda[d] and Fabrizio Ruffini[e]

*Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy*

Keywords: Fuzzy Regression Trees, Regression Models, Explainable Artificial Intelligence, Approximation Functions.

Abstract: Fuzzy Regression Trees (FRTs) are widely acknowledged as highly interpretable ML models, capable of dealing with noise and/or uncertainty thanks to the adoption of fuzziness. The accuracy of FRTs, however, strongly depends on the polynomial function adopted in the leaf nodes. Indeed, their modelling capability increases with the order of the polynomial, even if at the cost of greater complexity and reduced interpretability. In this paper we introduce the concept of Heterogeneous FRT: the order of the polynomial function is selected on each leaf node and can lead either to a zero-order or a first-order approximation. In our experimental assessment, the percentage of the two approximation orders is varied to cover the whole spectrum from pure zero-order to pure first-order FRTs, thus allowing an in-depth analysis of the trade-off between accuracy and interpretability. We present and discuss the results in terms of accuracy and interpretability obtained by the corresponding FRTs on nine benchmark datasets.

## 1 INTRODUCTION

In recent years, systems based on Artificial Intelligence (AI) and Machine Learning (ML) are rapidly changing the way new services are conceived and developed in the public and private sectors. The impact of the current applications is so significant that it is reflected, almost daily, by a worldwide media coverage discussing both the promises and the risks of an AI-powered society (Cath et al., 2018), (Fontes et al., 2022), (Leikas et al., 2022). The discussions mainly focus on how different stakeholders can trust the decision of AI: the need is to avoid violations of what are perceived as fundamental human rights, while increasing the welfare of the society as a whole. One component of trust is the capability to understand how an AI system works, so as to be able to understand and justify its outcome. This is of paramount importance in specific sectors such as health, defence, finance, and law, where the discussion is more intense given the high stakes involved. Thus, since recent years, legislators have started to take into account the topic of the trustworthiness of AI-systems, resulting in recommendations (for example, the "Ethics Guidelines for Trustworthy AI" (High-Level Expert Group on AI, 2019)) and in law proposals (as in the recent European "Artificial Intelligence Act" proposal (AIA, 2021)).

Explainable AI (XAI) is a research field focused on devising AI-systems understandable to the different stakeholders involved. Following the terminology in (Barredo Arrieta et al., 2020), there are two different strategies for achieving explainability: i) post-hoc strategies, which aim to describe a posteriori how an AI system works, and ii) ante-hoc strategies, which directly design models that are inherently explainable. Post-hoc explanations are typically applied to Neural Networks (NNs) and ensemble methods. Such families of models are generally referred to as *opaque* or "*black-box*", as opposed to *transparent* models, whose operation is inherently understandable for a human (Barredo Arrieta et al., 2020). Decision Trees (DTs), Regression Trees (RTs) and rule-based systems (RBSs) are considered among highly transparent and interpretable models.

Generally speaking, a distinction should be made between global and local interpretability. Global interpretability refers to the structure of the model: for example, in rule-based and tree-based models, the higher the number of rules and nodes, respectively,

[a] https://orcid.org/0000-0002-9984-1904
[b] https://orcid.org/0000-0003-4510-1350
[c] https://orcid.org/0000-0002-5895-876X
[d] https://orcid.org/0000-0002-0482-5048
[e] https://orcid.org/0000-0001-6328-4360



Bárcena, J., Ducange, P., Gallo, R., Marcelloni, F., Renda, A. and Ruffini, F.
Experimental Assessment of Heterogeneous Fuzzy Regression Trees.
DOI: 10.5220/0012231000003595
In *Proceedings of the 15th International Joint Conference on Computational Intelligence (IJCCI 2023)*, pages 376-384
ISBN: 978-989-758-674-3; ISSN: 2184-3236
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

the lower the global interpretability. Local interpretability, on the other hand, refers to the explanation of a specific decision taken by a model for any single input instance and is indeed related to the inference process. Local interpretability of DTs and RTs is high, in general, because they can be transformed into a set of "if-then" rules, which are consistent with a reasoning paradigm familiar to human beings. Specifically, the lower the number of antecedent conditions and parameters in the consequent of the activated rule, the higher the local interpretability for the specific decision.

In this paper, we focus on regression tasks and investigate the adoption of a highly interpretable tree-based ML model, namely an RT. Leaf nodes of RTs are characterized by an approximation polynomial function defined over the input variables. Different polynomial functions have been used in the literature. For instance M5 (Quinlan, 1992) employs first-order polynomials using the overall set of input variables. CART (Breiman et al., 1984), which is one of the most known algorithms for generating RTs, uses zero-order polynomials, leading to simpler and usually more robust models.

When first-order polynomial functions are adopted, two main approaches can be used (Bertsimas et al., 2021): the most straightforward approach consists in first growing the tree assuming constant predictions and then estimating a linear model in the leaves. On the one hand this strategy avoids the cost of repeatedly fitting linear models during training; on the other hand, decoupling the growing and the leaves regression estimate steps results in trees which are typically larger than what is actually needed. A popular alternative approach (Chaudhuri et al., 1995; Torsten Hothorn and Zeileis, 2006) relies on hypothesis testing to choose the variables for the splits: albeit computationally efficient, it is shown to produce trees with a limited generalization capability (Bertsimas et al., 2021). Some recent works (Bertsimas and Dunn, 2017; Dunn, 2018; Bertsimas and Dunn, 2019) have discussed the concept of Optimal Regression Tree (ORT), conceived to pursue greater predictive power. ORTs utilize mixed-integer optimization and local heuristic methods to find near-optimal trees both for axes-aligned splits and for splits with generic hyperplanes (i.e., not necessarily aligned with the axes). In the latter case, evidently, the interpretability is reduced.

In general, zero-order polynomials are simpler and easier to interpret compared to higher order ones. When using a higher order polynomial (typically first-order) in leaf nodes, modelling capacity increases at the cost of increased complexity and, in turn, decreased interpretability.

An extensive literature has focused on the integration of fuzzy set theory with decision and regression trees (Suárez and Lutsko, 1999), (Chen et al., 2009), (Segatori et al., 2018), (Cózar et al., 2018), (Renda et al., 2021), (Bechini et al., 2022). The adoption of fuzziness is typically meant to bring a higher accuracy in scenarios characterized by vagueness and/or noise.

In this paper, we report on an in-depth analysis of the performance of a Fuzzy Regression Tree (FRT) on several benchmark datasets exploiting the novel concept of *heterogeneity*: in a heterogeneous FRT, the order of the polynomial to be used as approximation is selected on each leaf node. In this paper, we focus on zero-order and first-order approximations and therefore some leaf nodes will have zero-order regression models and some others will have first-order regression models. By exploring the whole spectrum (from purely zero-order to purely first-order approximations) we investigate several trade-offs between accuracy and interpretability and derive insights about the viability of intermediate, heterogeneous, solutions.

This work stems from a previous work presented in (Bechini et al., 2022), where we compared the performance of pure FRTs with only zero-order approximators against pure FRTs with only first-order approximators. Indeed, this paper entails the following contributions:

- we introduce the concept of *heterogeneous* FRT, allowing for leaves with polynomial models of different orders;
- we assess the performance of the proposed approach on several benchmark regression datasets;
- we investigate the trade-off between accuracy and interpretability for different degrees of heterogeneity.

The paper is organized as follows: in Section 2 we provide some background on the FRT model adopted in this paper. In Section 3 we present the concept of heterogeneous FRT. In Section 4 we report the experimental setup and results. Finally in Section 5 we draw our conclusions.

## 2 BACKGROUND: FUZZY REGRESSION TREE

Let $\mathbf{X} = \{X_1, \ldots, X_F\}$ be the set of input variables. A regression tree is a directed acyclic graph, where each internal (non-leaf) node represents a test on an input variable and each leaf is characterized by a regression model. Each path from the root to one leaf

corresponds to a sequence of tests. The format of the tests depends on the type of the input variables. In the case of numerical variables, tests are in the form $X_f > x_{f,s}$ and $X_f \leq x_{f,s}$, where $x_{f,s} \in \mathbb{R}$. This type of test results in binary trees, in which each internal node has at most two child nodes. In the case of categorical variables, tests are in the form $X_f \subseteq L_{f,s}$, where $L_{f,s}$ is a subset of possible categorical values for $X_f$; as a consequence, each test may result in more than two branches, thus originating the so-called multi-way trees.

Our work stems from the proposals for building FRTs presented in (Cózar et al., 2018) and revisited in (Bechini et al., 2022). We assume that each real input variable $X_f$ is partitioned by using $T_f$ fuzzy sets. Let $\mathbf{P}_f = \{B_{f,1}, \ldots, B_{f,T_f}\}$ be the partition of input variable $X_f$. The tests in the internal nodes use these fuzzy sets in the form of "$X_f$ is $B_{f,j}$". Since fuzzy sets generally overlap, an input instance may activate more than one leaf node. We employ multi-way trees and use all the $T_f$ fuzzy sets for the tests on input variable $X_f$ in one node, thus generating $T_f$ branches.

In the case of a zero-order polynomial regression model, the value $\phi^{(K)}(\mathbf{X})$ assigned to each leaf node $LN^{(K)}$ is a constant, which is computed as a weighted average of the output values $y_i$ of all the instances in the training set that activate such leaf node, where the weight $w_{LN^{(K)}}$ is the strength of activation of the path $R^{(K)}$ from the root to the leaf node $LN^{(K)}$. More formally, given an input pattern $\mathbf{x}_i$ corresponding to the output value $y_i$, value $\phi^{(K)}(\mathbf{X})$ is computed as:

$$\phi^{(K)}(\mathbf{X}) = c^{(K)} = \frac{\sum_{(\mathbf{x}_i,y_i)|w_{LN^{(K)}}(\mathbf{x}_i)>0} \left(y_i \cdot w_{LN^{(K)}}(\mathbf{x}_i)\right)}{\sum_{(\mathbf{x}_i,y_i)|w_{LN^{(K)}}(\mathbf{x}_i)>0} \left(w_{LN^{(K)}}(\mathbf{x}_i)\right)} \tag{1}$$

where

$$w_{LN^{(K)}}(\mathbf{x}_i) = \prod_{k=1}^{K} \mu_{f^{(k)}}(x_{i,f^{(k)}}) \tag{2}$$

The term $\mu_{f^{(k)}}(x_{i,f^{(k)}})$ is the membership degree of $x_{i,f^{(k)}}$ to the fuzzy set $B_{f^{(k)},j}$ of the partition of each input variable $X_{f^{(k)}}$ chosen in each node $N^{(k)}$ in the path from the root ($k = 1$) to the leaf node $LN^{(K)}$ ($k = K$).

A first-order polynomial regression model employs a linear model in any leaf node. The model is defined as follows:

$$\phi^{(K)}(\mathbf{X}) = \gamma_0^{(K)} + \sum_{f=1}^{F} \gamma_f^{(K)} \cdot X_f \tag{3}$$

The coefficients $\Gamma^{(K)} = \left\{\gamma_0^{(K)}, \gamma_1^{(K)}, \ldots, \gamma_F^{(K)}\right\}$ can be estimated by applying a local weighted least-squared method. Specifically, in the estimation of the parameters, each training sample $(\mathbf{x}_i, y_i)$ with a membership value greater than 0 to the specific leaf is weighted by its strength of activation of the rule. Notably, for any given rule, the linear regression model considers the whole set of $F$ input variables, even if typically only a subset of them appears in the antecedent part.

## 2.1 Partition Fuzzy Gain, Fuzzy Variance and Fuzzy Mean

In regression problems, the sequence of tests aims to partition the input space into subspaces that contain subsets of the training set with output values as close as possible to each other. In the learning phase, the choice of the input variable to be used in a decision node is generally performed based on the variance of the output values.

Similar to (Cózar et al., 2018), in our FRT the splitting criterion is based on the Partition Fuzzy Gain (*PFGain*) index, which in turn hinges on the concept of Fuzzy Variance. Formally, let $N^{(k)}$ be a generic node in FRT. The quantity $w_{N^{(k)}}(\mathbf{x}_i) = \prod_{t=1}^{k} \mu_{B_{f^{(k)},j}}(x_{i,f^{(k)}})$ is the strength of activation of instance $(\mathbf{x}_i, y_i) \in TR$ to node $N^{(k)}$ computed along the path $R^{(k)}$ from the root to $N^{(k)}$, where $TR = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Z, y_Z)\}$ is the training set of $Z$ instances. In the root, $w_{N^1}(\mathbf{x}_i) = 1$ for all the instances in $TR$. Let $S^{N^{(k)}} = \{(\mathbf{x}_i, y_i) \in TR \mid w_{N^{(k)}}(\mathbf{x}_i) > 0\}$ be the set of instances with non-null strength of activation to node $N^{(k)}$. In the following, we first introduce the fuzzy mean and the fuzzy variance for the instances in $S^{N^{(k)}}$. Then, we define the fuzzy mean and the fuzzy variance of the instances in the support of a generic fuzzy set $B_{f^{(k)},j}$ when the instances in $S^{N^{(k)}}$ are partitioned by using $\mathbf{P}_{f^{(k)}}$. Finally, we introduce the definition of *PFGain*.

The *Fuzzy Mean* $FM^{N^{(k)}}$ of node $N^{(k)}$ is defined as the mean of the output values $y_i$ of the instances $(\mathbf{x}_i, y_i)$ in $S^{N^{(k)}}$, weighted by the strength of activation $w_{N^{(k)}}(\mathbf{x}_i)$:

$$FM^{N^{(k)}} = \frac{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}} \left(y_i \cdot w_{N^{(k)}}(\mathbf{x}_i)\right)}{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}} \left(w_{N^{(k)}}(\mathbf{x}_i)\right)} \tag{4}$$

The *Fuzzy Variance* $FVar^{N^{(k)}}$ of node $N^{(k)}$ is defined as follows:

$$FVar^{N^{(k)}} = \frac{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}} \left(y_i - FM^{N^{(k)}}\right)^2 \cdot \left(w_{N^{(k)}}(\mathbf{x}_i)\right)^2}{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}} \left(w_{N^{(k)}}(\mathbf{x}_i)\right)^2} \tag{5}$$

Let $S^{N^{(k)}}_{f,1},...,S^{N^{(k)}}_{f,T_F}$ be the subsets of points in $S^{N^{(k)}}$, contained in the supports of the fuzzy sets $B_{f^{(k)},1},...,B_{f^{(k)},T_f}$ of the partition $P_f$ tested for splitting node $N^{(k)}$.

The fuzzy mean $FM^{N^{(k)}}(B_{f^{(k)},j})$ of the output values computed for the instances of the support of fuzzy set $B_{f^{(k)},j}$ in the node $N^{(k)}$ is defined as the mean of the $y_i$ weighted by the product between the strength of activation of $x_{i,f^{(k)}}$ to the node $N^{(k)}$ and the membership degree of $x_{i,f^{(k)}}$ to $B_{f^{(k)},j}$:

$$FM^{N^{(k)}}(B_{f^{(k)},j}) = \frac{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}_{f^{(k)},j}} \left( y_i \cdot w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{f^{(k)},j}}(x_{i,f^{(k)}}) \right)}{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}_{f^{(k)},j}} \left( w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{f^{(k)},j}}(x_{i,f^{(k)}}) \right)} \tag{6}$$

The fuzzy variance $FVar^{N^{(k)}}(B_{f^{(k)},j})$ of the output values computed for the instances of the support of fuzzy set $B_{f^{(k)},j}$ in the node $N^{(k)}$ is defined as follows:

$$FVar^{N^{(k)}}(B_{f^{(k)},j}) = \frac{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}_{f^{(k)},j}} \left( y_i - FM^{N^{(k)}}(B_{f^{(k)},j}) \right)^2 \cdot \left( w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{f^{(k)},j}}(\mathbf{x}_{i,f^{(k)}}) \right)^2}{\sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}_{f,j}} \left( w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{f^{(k)},j}}(\mathbf{x}_{i,f^{(k)}}) \right)^2} \tag{7}$$

Finally, let $WS^{N^{(k)}}(B_{f^{(k)},j})$ be the following quantity:

$$WS^{N^{(k)}}(B_{f^{(k)},j}) = \sum_{(\mathbf{x}_i,y_i)\in S^{N^{(k)}}_{f,j}} w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{f^{(k)},j}}. \tag{8}$$

The *Partition Fuzzy Gain* $PFGain^{N^{(k)}}(P^{(k)}_f)$ obtained by adopting the fuzzy partition $P^{(k)}_f$ over the input variable $X^{(k)}_f$ is defined as follows:

$$PFGain^{N^{(k)}}(P^{(k)}_f) = FVar^{N^{(k)}} - \sum_{j=1}^{T_f} FVar^{N^{(k)}}(B_{f^{(k)},j}) \cdot W^{N^{(k)}}(B_{f^{(k)},j}) \tag{9}$$

where:

$$W^{N^{(k)}}(B_{f^{(k)},j}) = \frac{WS^{N^{(k)}}(B_{f^{(k)},j})}{\sum_{j=1}^{T_f} WS^{N^{(k)}}(B_{f^{(k)},j})} \tag{10}$$

Let $X_{\hat{f}}$ be the input variable with the highest $PFGain^{N^{(k)}}$. Fuzzy sets $B_{\hat{f},j}$ are used to split the instances in node $N^{(k)}$ into $T_{\hat{f}}$ child nodes $N^{(k+1)}_j$, $j = [1,...,T_{\hat{f}}]$. The strength of activation $w_{N^{(k+1)}_j}(\mathbf{x}_i)$ of a generic instance $\mathbf{x}_i$ to child node $N^{(k+1)}_j$ is computed as $w_{N^{(k+1)}_j}(\mathbf{x}_i) = w_{N^{(k)}}(\mathbf{x}_i) \cdot \mu_{B_{\hat{f},j}}(x_{i,\hat{f}})$.

## 2.2 Tree Construction and Inference Process

Let $N^{(k)}$ be a generic node at depth $k$. Let $Z^{N^{(k)}} = \{(\mathbf{x}_i,y_i) \in TR \mid w_{N^{(k)}}(\mathbf{x}_i) \geq 0.5^k\}$ be the set of instances that *strongly* activate the node, i.e., having a activation strength higher than 0.5 raised to the tree level at which $N^{(k)}$ is located (in other words, we are assuming that at each level of the path the instances belong to the corresponding fuzzy set with a membership degree higher than 0.5). In the proposed algorithm the following criteria are employed to stop the tree growth at a generic node $N^{(k)}$:

- when the cardinality of $Z^{N^{(k)}}$ is lower than a fraction (*min_samples_split*) of the cardinality of the training set $TR$;
- when the highest *PFGain* computed for $N^{(k)}$ is lower than a fixed threshold (*min_PFGain*);
- when the set of input variables available for splitting $N^{(k)}$ is empty.

The tree growing procedure terminates when there exist no nodes that can be considered for possible splitting. Once the tree has been generated, a regression model is assigned to each leaf node. In our proposal, we employ both zero-order and first-order polynomial regression models (see Equations 1 and 3).

The path $p_r$ from the root to a generic leaf node $LN^{(K)}$ at the $K^{th}$ level corresponds to the following rule $R_r$:

$$R_r: \textbf{IF } X_{r^{(2)}} \textit{ is } B_{r^{(2)},j_{r^{(2)}}} \textbf{ AND } \ldots \textbf{ AND } X_{r^{(K)}} \textit{ is } B_{r^{(K)},j_{r^{(K)}}} \textbf{ THEN } Y = \phi_r(\mathbf{X}) \tag{11}$$

where $X_{r^{(k)}}$ and $B_{r^{(k)},j_{r^{(k)}}}$ are, respectively, the input variable and the fuzzy set of the corresponding partition which allow reaching the node at the $k^{th}$ level of path $p_r$ and contribute to the strength of activation for this node (we recall that $(k = 1)$ identifies the root node).

Given an input pattern $\hat{\mathbf{x}}$, the inference process generates an output based on the maximum matching strategy: only the rule with the highest strength of activation is used for estimating the output value.

The strength of activation of the rule $R_r$ is computed as:

$$w_r(\hat{\mathbf{x}}) = \prod_{k=1}^{K} \mu_{B_{r^{(k)},j_{r^{(k)}}}}(\hat{x}_{r^{(k)}}) \tag{12}$$

It has been shown that the adoption of a maximum matching approach does not particularly degrade the modelling power compared to the weighted average

strategy, yet ensuring a higher level of interpretability (Bechini et al., 2022).

The use of the product as a T-norm operator (see Eq. 2) for the computation of the strength of activation has an obvious implication on the inference process: since the terms $\mu_{B_{r(k),j_{r(k)}}}(\hat{x}_{r(k)})$ are in the range $[0,1]$, the maximum matching approach in general prioritizes short rules, i.e., those associated with leaf nodes *closer* to the root of the FRT. To compensate for this phenomenon, we consider the normalized strength of activation $\tilde{w}_r(\hat{\mathbf{x}})$, which is defined as follows:

$$\tilde{w}_r(\hat{\mathbf{x}}) = \frac{w_r(\hat{\mathbf{x}})}{\bar{w}_r(TR)} \tag{13}$$

where $\bar{w}_r(TR)$ is the average strength of activation for all instances $\mathbf{x}_i$ in the training set with $w_r(\mathbf{x}_i) > 0$.

# 3 HETEROGENEOUS FRT

The implementation of linear regression models in the leaf nodes enhances the modelling capability of the FRT but it is not without drawbacks.

First of all, the adoption of a linear model reduces both *local* and *global* interpretability. From a local perspective, which refers to how a prediction is carried out during the inference process, in the linear model the effect of each input variable on the output value is expressed by the corresponding coefficient; the zero-order model, on the other hand, provides the output value directly. From a global perspective, which refers to the structural properties of the model, the number of parameters increases with the order of the polynomial: in general, the complexity of the model can be considered as a proxy for its interpretability.

Second, a more complex model is more prone to the phenomenon of overfitting, whereby an increased *modelling* capability is not matched by an increased *generalization* capability.

In this paper, we introduce the concept of *heterogeneous* FRT, in which the choice of the order of the model to be used is evaluated on each individual leaf node. The basic idea is to assess on each leaf node whether the first-order model is needed or whether the zero-order model can be used.

The criterion for the model selection is based on the concept of Fuzzy Variance, as reported in Eq. 5. In particular, once the tree structure is generated, the *FVar* associated with all the leaves is evaluated and the selection of the model is based on a threshold $th_{order} \in [0,100]$: specifically, a percentage of the leaves equal to $th_{order}$, those with lower *FVar* values, employ zero-order models, while the others employ first-order models. Indeed, it is reasonable to assume that a zero-order model is sufficient where the value of *FVar* is lower, whereas a more complex model is needed where the *FVar* is higher.

# 4 EXPERIMENTAL ANALYSIS

This section reports on the experimental setup and results obtained with our heterogenouos FRT on several regression datasets. We first describe the datasets and the model parameters. Then, we discuss the numerical results.

## 4.1 Experimental Setup

The regression datasets employed in our experimental analysis are publicly available within the Keel (Alcalá-Fdez et al., 2011) (Anacalt, Elevators, House, Weather Izmir, Treasury, Mortgage) and Torgo's[1] (Puma8NH, California, Kinematics) dataset repositories. The number of samples and input variables for each of the datasets are reported in Table 1.

Table 1: Datasets description.

| Dataset | # Input Variables | # Samples |
|---|---|---|
| Puma8NH (PU) | 8 | 8192 |
| ANACALT (AN) | 7 | 4052 |
| Elevators (EL) | 18 | 16599 |
| House (HO) | 16 | 22784 |
| Weather Izmir (WI) | 9 | 1461 |
| Treasury (TR) | 15 | 1049 |
| Mortgage (MO) | 15 | 1049 |
| California (CA) | 8 | 20460 |
| Kinematics (KI) | 8 | 8192 |

In our preliminary experimental assessment, the following configuration parameters are considered for FRT induction:

- a strong uniform fuzzy partition based on five triangular fuzzy sets is employed on each input attribute. The five fuzzy sets can be labelled with the following linguistic terms: *VeryLow*, *Low*, *Medium*, *High* and *VeryHigh*.
- $min\_samples\_split = 0.1$;
- $min\_PFGain = 0.0001$.

Furthermore, a robust scaling (using 2.5 and 97.5 percentiles) is applied to the input variables to remove outliers and clip the distribution in the range $[0,1]$.

[1] https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html

In order to assess the performance of the proposed heterogeneous FRT and investigate the trade-off between accuracy and interpretability we carried out an experimental analysis by considering the following values for the threshold $th_{order}$: $\{0,5,10,20,40,60,80,100\}$. We recall that $th_{order} = i$ implies that $i\%$ of the leaves with the lowest variance use a zero-order model and $(100-i)\%$ use a first-order model. For $th_{order} = 0$ and $th_{order} = 100$ we have purely first-order and purely zero-order FRTs, respectively.

The predictive capability of the heterogeneous FRTs is evaluated through the *Mean Squared Error* (MSE):

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (y_i - \hat{y}_i)^2 \tag{14}$$

where $N_{test}$ is the number of samples considered for the evaluation, $y_i$ and $\hat{y}_i$ are the ground truth value and the predicted value associated with the $i$-th instance of the test set, respectively. Results are evaluated in terms of average values over 5-fold cross-validation: at each iteration of the cross-validation, the same split is used for the different values of $th_{order}$.

## 4.2 Experimental Results

Table 2 reports the average results obtained by the heterogeneous FRTs for different values of the threshold $th_{order}$. Best values of MSE, averaged over 5-fold cross-validation, are highlighted in bold.

It is worth noticing that, for any dataset, the overall number of leaves (evaluated as the sum of the number of leaves employing zero- and first-order models) is constant, regardless of the $th_{order}$ value. This is expected, since the choice of the model order takes place after the tree growing procedure and has no impact on the number of leaves. Furthermore, it can be observed that the pre-pruning strategies, implemented by the stop conditions based on *min_samples_split* and *min_PFGain*, allow obtaining FRTs with a limited number of leaves (generally lower than 100). Obtaining trees with a relatively low number of leaves, which corresponds to the number of rules, is crucial for the *global* interpretability of the models.

As for the accuracy of the FRTs (measured in terms of MSE on the test sets), it can be observed that FRTs featuring only first-order regression models in the leaves ($th_{order} = 0$) always outperform FRTs featuring only zero-order models ($th_{order} = 100$). First-order models entail both a higher modelling capability (lower MSE values on the training set) and generalization capability (lower MSE values on the test set) compared to the zero-order ones. Furthermore, the quality of the predictions of FRTs is comparable to that measured in previous relevant works (Antonelli et al., 2011), (Bechini et al., 2022). It can be observed, however, that first-order models can be prone to the overfitting phenomenon: on Mortgage and Treasury datasets, in particular, the MSE on the test set is higher than the MSE on the training set by a factor around 2, with $th_{order} = 0$. Such factor decreases with increasing values of $th_{order}$. This phenomenon is probably explained by the low numerosity of the two datasets.

Expectedly, the heterogeneity of the models employed in the leaves allows for intermediate results between those achieved by purely zero-order and purely first-order FRTs. The only exception occurs for the Puma8NH dataset, where modelling few leaves (2 or 3 out of around 26) with zero-order models entails a minimal gain in accuracy (MSE on the test set decreases from 10.16 to 10.15). Otherwise, in general, heterogeneous FRTs perform gradually worse as $th_{order}$ increases. It is worth pointing out, however, that adopting a zero-order model in a small fraction of leaves with the lowest variance ($th_{order} \in [5,10,20]$) does not entail a significant degradation of test accuracy. At the same time, the resulting models are simpler than in the case of $th_{order} = 0$. To quantify the gain in complexity and thus in interpretability, we extend the definition of FRT complexity ($C_{FRT}$) adopted in (Bechini et al., 2022):

$$C_{FRT} = IN + LN_0 + LN_1 \times (F+1) \tag{15}$$

where $IN$, $LN_0$ and $LN_1$ represent the number of internal nodes, the number of leaves implementing a zero-order model and the number of leaves implementing a first order model, respectively. In other words, the formulation of $C_{FRT}$ captures the overall number of parameters of an FRT, noting that each zero-order model has only one parameter (i.e., the constant value) whereas a first-order model has $F+1$ parameters (i.e., the vector of coefficients $\Gamma$ of the linear model).

To better illustrate the trade-off between accuracy and complexity (and indeed interpretability) of heterogeneous FRTs, in Fig. 1 we report the values of MSE and $C_{FRT}$ for each dataset and for each value of $th_{order}$. Plots confirm the observation that MSE values tend to increase with $th_{order}$: the performance degradation is more evident for high value of the threshold, but is less evident for lower values.

By the analysis of the experimental results, we can conclude that the use of zero-order regression models in the 10-20% of leaf nodes characterised by the lowest fuzzy variance does not affect considerably the performance of the FRTs, but produces considerably

Table 2: Experimental results: average MSE results obtained by varying the $th_{order}$ parameter, together with the respective number of leaves implementing a zero-order and first-order regression model. Best values of MSE are highlighted in bold.

| | train | *test* | 0-ord. | 1-ord. | train | *test* | 0-ord. | 1-ord. | train | *test* | 0-ord. | 1-ord. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $th_{order}$ | **Puma8NH** | | | | **Anacalt** ($\times 10^{-2}$) | | | | **Elevators** ($\times 10^{-6}$) | | | |
| **0** | 9.80 | 10.16 | 0.0 | 25.8 | 1.23 | **1.41** | 0.0 | 13.8 | 6.73 | **7.24** | 0.0 | 75.8 |
| **5** | 9.81 | **10.15** | 2.0 | 23.8 | 1.23 | **1.41** | 3.0 | 10.8 | 6.74 | 7.25 | 4.4 | 71.4 |
| **10** | 9.82 | **10.15** | 3.0 | 22.8 | 1.23 | **1.41** | 3.0 | 10.8 | 6.77 | 7.28 | 8.2 | 67.6 |
| **20** | 9.98 | 10.30 | 6.0 | 19.8 | 1.23 | **1.41** | 3.4 | 10.4 | 6.94 | 7.54 | 16.0 | 59.8 |
| **40** | 10.30 | 10.57 | 11.2 | 14.6 | 1.84 | 2.02 | 6.2 | 7.6 | 7.41 | 8.01 | 31.0 | 44.8 |
| **60** | 10.92 | 11.13 | 16.4 | 9.4 | 2.34 | 2.50 | 8.8 | 5.0 | 8.51 | 9.19 | 46.2 | 29.6 |
| **80** | 11.72 | 11.89 | 21.6 | 4.2 | 3.08 | 3.35 | 11.6 | 2.2 | 10.85 | 11.46 | 61.2 | 14.6 |
| **100** | 12.62 | 12.64 | 25.8 | 0.0 | 6.58 | 6.76 | 13.8 | 0.0 | 24.16 | 24.39 | 75.8 | 0.0 |
| $th_{order}$ | **House** ($\times 10^{9}$) | | | | **Weather Izimir** | | | | **Treasury** ($\times 10^{-2}$) | | | |
| **0** | 1.33 | **1.46** | 0.0 | 91.8 | 1.00 | **1.68** | 0.0 | 93.8 | 2.51 | **5.40** | 0.0 | 54.0 |
| **5** | 1.33 | **1.46** | 5.0 | 86.8 | 1.00 | 1.70 | 5.2 | 88.6 | 2.51 | 5.64 | 3.2 | 50.8 |
| **10** | 1.33 | **1.46** | 9.6 | 82.2 | 1.01 | 1.72 | 9.8 | 84.0 | 2.58 | 5.69 | 6.0 | 48.0 |
| **20** | 1.34 | 1.47 | 18.8 | 73.0 | 1.09 | 1.84 | 19.4 | 74.4 | 3.01 | 6.26 | 11.4 | 42.6 |
| **40** | 1.38 | 1.50 | 37.4 | 54.4 | 1.62 | 2.54 | 38.2 | 55.6 | 4.81 | 8.06 | 22.2 | 31.8 |
| **60** | 1.47 | 1.59 | 55.6 | 36.2 | 2.81 | 4.03 | 56.8 | 37.0 | 8.25 | 12.61 | 32.8 | 21.2 |
| **80** | 1.70 | 1.78 | 74.2 | 17.6 | 4.75 | 5.85 | 75.6 | 18.2 | 11.98 | 15.18 | 43.6 | 10.4 |
| **100** | 2.06 | 2.08 | 91.8 | 0.0 | 7.11 | 8.15 | 93.8 | 0.0 | 34.42 | 36.62 | 54.0 | 0.0 |
| $th_{order}$ | **Mortgage** ($\times 10^{-2}$) | | | | **California** ($\times 10^{9}$) | | | | **Kinematics** ($\times 10^{-2}$) | | | |
| **0** | 0.95 | **1.92** | 0.0 | 100.4 | 3.34 | **3.45** | 0.0 | 87.8 | 3.67 | **3.81** | 0.0 | 25.0 |
| **5** | 0.95 | **1.92** | 5.8 | 94.6 | 3.35 | **3.45** | 4.8 | 83.0 | 3.69 | 3.82 | 2.0 | 23.0 |
| **10** | 0.95 | **1.92** | 10.4 | 90.0 | 3.36 | 3.47 | 9.0 | 78.8 | 3.71 | 3.84 | 3.0 | 22.0 |
| **20** | 1.02 | 1.98 | 20.6 | 79.8 | 3.42 | 3.53 | 18.0 | 69.8 | 3.77 | 3.90 | 6.0 | 19.0 |
| **40** | 2.13 | 3.24 | 40.8 | 59.6 | 3.69 | 3.79 | 35.6 | 52.2 | 3.95 | 4.07 | 11.0 | 14.0 |
| **60** | 3.68 | 4.63 | 60.8 | 39.6 | 4.29 | 4.42 | 53.4 | 34.4 | 4.26 | 4.38 | 16.0 | 9.0 |
| **80** | 6.58 | 9.01 | 81.0 | 19.4 | 5.21 | 5.32 | 71.0 | 16.8 | 4.45 | 4.53 | 21.0 | 4.0 |
| **100** | 18.48 | 20.33 | 100.4 | 0.0 | 5.88 | 5.95 | 87.8 | 0.0 | 4.59 | 4.62 | 25.0 | 0.0 |

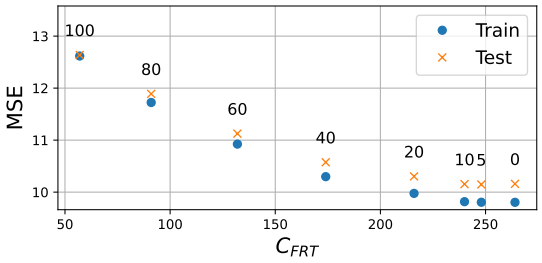
advantages in terms of both global and local interpretability.

## 5 CONCLUSION

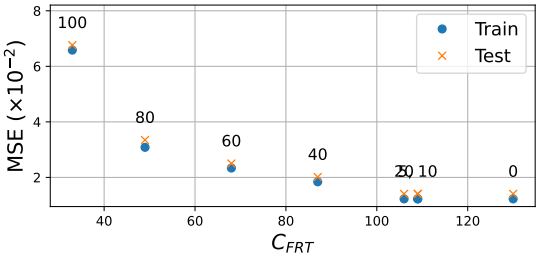
In this paper, we have introduced the concept of heterogeneity in Fuzzy Regression Tree (FRT), allowing for different polynomial approximators (i.e., either zero- or first-order models) in different leaf nodes of an FRT. The model selection criterion is based on the concept of Fuzzy Variance of the leaf nodes. We investigate the trade-off between interpretability and accuracy (expressed as Mean Squared Error) for different degrees of heterogeneity on nine benchmark regression datasets. The results showed that, in general, the heterogeneous FRTs achieve intermediate performance between the pure zero-order and the pure first-order FRTs. In detail, first-order models entail higher predictive capability (i.e., lower MSE values) compared to zero-order ones, but this comes at a cost of an increased complexity and reduced interpretability. Interestingly, heterogeneous FRTs with a small quota of leaves employing zero-order models (i.e. from 5% to 20%) provide a gain in interpretability compared to purely first-order FRTs, without significant loss in terms of MSE. In conclusion, the proposed heterogeneous FRT has proven its effectiveness in scenarios where, given a performance constraint, it is necessary to optimize the model's explainability by reducing the number of model parameters. Future works will investigate the sensitivity of Heterogenous FRTs with respect to its main hyperparameters and will comprise comparative experiments with other classical and state-of-art ML approaches, in terms of accuracy and explainability.

## ACKNOWLEDGEMENTS

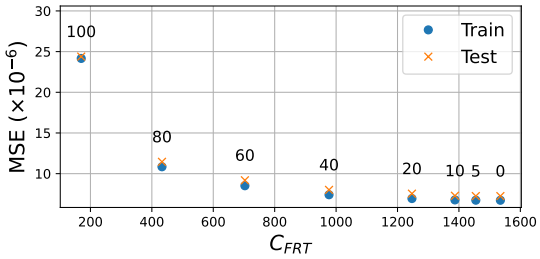
This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" and the PNRR "Tuscany Health Ecosystem" (THE) (Ecosistemi dell'Innovazione) - Spoke 6 - Precision Medicine & Personalized Healthcare (CUP I53C22000780001) under the NextGeneration EU programme, and by the



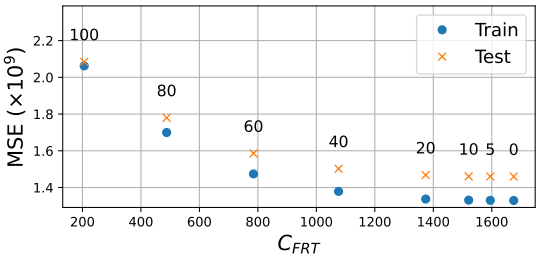

(a) Puma8NH



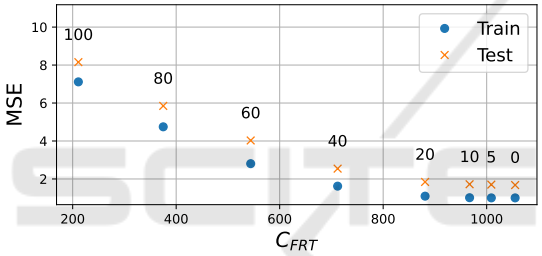

(b) Anacalt



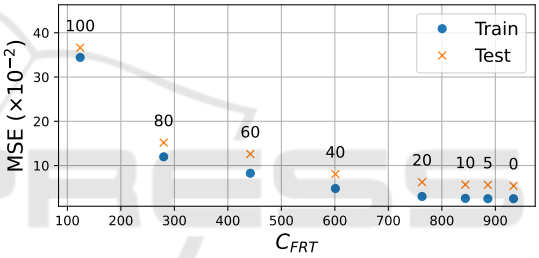

(c) Elevators



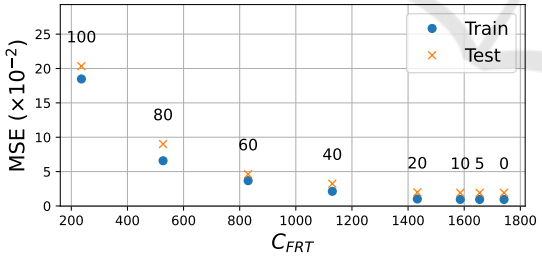

(d) House



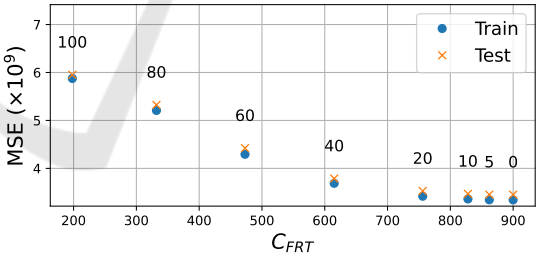

(e) Weather Izimir



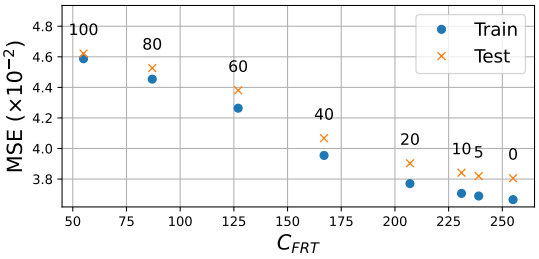

(f) Treasury




(g) Mortgage




(h) California




(i) Kinematics

Figure 1: Plots of the average MSE on the training and test sets and average complexity (i.e., overall number of parameters) of the FRTs for different values of $th_{model}$, as reported in the annotations within the figures.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence).

## REFERENCES

(2021). Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

Antonelli, M., Ducange, P., and Marcelloni, F. (2011). Genetic training instance selection in multiobjective evolutionary fuzzy systems: A coevolutionary approach. *IEEE Transactions on fuzzy systems*, 20(2):276–290.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bechini, A., Bárcena, J. L. C., Ducange, P., Marcelloni, F., and Renda, A. (2022). Increasing Accuracy and Explainability in Fuzzy Regression Trees: An Experimental Analysis. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106:1039–1082.

Bertsimas, D. and Dunn, J. (2019). *Machine learning under a modern optimization lens*. Dynamic Ideas LLC Charlestown, MA.

Bertsimas, D., Dunn, J., and Wang, Y. (2021). Near-optimal nonlinear regression trees. *Operations Research Letters*, 49(2):201–206.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Cart. *Classification and regression trees*.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24:505–528.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica*, pages 641–666.

Chen, Y.-l., Wang, T., Wang, B.-s., and Li, Z.-j. (2009). A survey of fuzzy decision tree classifier. *Fuzzy Inf. Eng.*, 1(2):149–159.

Cózar, J., Marcelloni, F., Gámez, J. A., and de la Ossa, L. (2018). Building efficient fuzzy regression trees for large scale and high dimensional problems. *Journal of Big Data*, 5(1):1–25.

Dunn, J. W. (2018). *Optimal trees for prediction and prescription*. PhD thesis, Massachusetts Institute of Technology.

Fontes, C., Hohma, E., Corrigan, C. C., and Lütge, C. (2022). Ai-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, 71:102137.

High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy ai. Report, European Commission, Brussels.

Leikas, J., Johri, A., Latvanen, M., Wessberg, N., and Hahto, A. (2022). Governing ethical ai transformation: A case study of auroraai. *Frontiers in Artificial Intelligence*, 5:13.

Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific.

Renda, A., Ducange, P., Gallo, G., and Marcelloni, F. (2021). XAI Models for Quality of Experience Prediction in Wireless Networks. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

Segatori, A., Marcelloni, F., and Pedrycz, W. (2018). On distributed fuzzy decision trees for big data. *IEEE Transactions on Fuzzy Systems*, 26(1):174–192.

Suárez, A. and Lutsko, J. F. (1999). Globally optimal fuzzy decision trees for classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1297–1311.

Torsten Hothorn, K. H. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.