TREE-STRUCTURED SMOOTH TRANSITION REGRESSION MODELS

JOEL CORRÊA DA ROSA, ALVARO VEIGA, AND MARCELO C. MEDEIROS

ABSTRACT. This paper introduces a tree-based model that combines aspects of CART (Classification and Regression Trees) and STR (Smooth Transition Regression). The model is called the Smooth Transition Regression Tree (STR-Tree). The main idea relies on specifying a parametric nonlinear model through a tree-growing procedure. The resulting model can be analyzed as a smooth transition regression with multiple regimes. Decisions about splits are entirely based on a sequence of Lagrange Multiplier (LM) tests of hypotheses. An alternative specification strategy based on a 10-fold cross-validation is also discussed and a Monte Carlo experiment is carried out to evaluate the performance of the proposed methodology in comparison with standard techniques. The STR-Tree model outperforms CART when the correct selection of the architecture of simulated trees is discussed. Furthermore, the LM test seems to be a promising alternative to 10-fold cross-validation. Function approximation is also analyzed. When put into proof with real and simulated datasets, the STR-Tree model has a superior predictive ability than CART.

KEYWORDS. Regression-trees, CART, smooth transition, nonlinear regression, modeling cycle.

ACKNOWLEDGMENTS. This work is partly supported by CNPq and FAPERJ. We wish to thank two anonymous referees for helpful comments.

1. INTRODUCTION

IN RECENT YEARS much attention has been devoted to nonlinear modeling. Techniques such as artificial neural networks, nonparametric regression and recursive partitioning methods are frequently used to approximate unknown functional forms (Murthy 1998, Hastie, Tibshirani, and Friedman 2001). This paper considers a nonlinear regression model that combines aspects of two well-known methodologies: Classification and Regression Trees (CART) discussed in Breiman, Friedman, Olshen, and Stone (1984) and the Smooth Transition Regression (STR) presented in Granger and Teräsvirta (1993). The proposed model is called the Smooth Transition Regression Tree (STR-Tree). The CART methodology represents a unification of all tree-based classification and prediction methods that have been developed since Morgan and Sonquist (1963). It transformed the regression tree models in an important nonparametric alternative to the classical methods of regression. Since then, the attractiveness of this methodology has motivated many authors to create hybrid modeling strategies that merge tree techniques with known statistical methods. See, for example, Segal (1992) in a context of longitudinal data analysis, Ahn (1996) for survival analysis, and Cooper (1998) for time series analysis. Other approaches can be found in Ciampi (1991), Crowley and Blanc (1993), and Denison, Mallik, and Smith (1998).

Allowing smooth splits on the tree nodes instead of sharp ones, we associate each tree architecture with a smooth transition regression model and thus it turns possible to formulate a splitting criteria that are entirely based on statistical tests of hypotheses. The Lagrange Multiplier (LM) test in the context presented by Luukkonen, Saikkonen, and Teräsvirta (1988) is adapted for deciding if a node should be split or not. The tree growing procedure is used as a tool for specifying a parametric model that can be analyzed either as STR model or as a fuzzy regression (Jajuga 1986). In the former case, we can obtain confidence intervals for the parameters estimates in the tree leaves and predicted values. In the regression-tree literature, the replacement of sharp splits by soft (of smooth) thresholds is not a new idea; see Chang and Pavlidis (1977), Jang (1994), Yuan and Shaw (1995), Janickow (1998), Suárez and Lutsko (1999), and Olaru and Wehenkel (2003). However, we contribute to regression-tree literature by proposing a coherent model building strategy fully based on statistical arguments. Our proposal is simple, easily implemented, and is not computer intensive. Furthermore, decisions based on statistical inference also lessen the importance of post-pruning techniques to reduce the model complexity and circumvent identification problems common in nonlinear regressions; see Medeiros, Teräsvirta, and Rech (2006) for a related discussion. An alternative specification strategy based on a

10-fold cross-validation is considered. An extension of the basic model to allow for the inclusion of categorical variables is discussed. A detailed Monte Carlo experiment is carried out to evaluate the performance of the proposed methodology in comparison with standard techniques. The STR-Tree model outperforms CART when the correct selection of the architecture of simulated trees is considered. Even when the true model is a regression-tree with sharp splits, the model building strategy proposed here selects the correct architecture in almost 100% of the cases. Furthermore, the LM test is less computer-intensive than 10-fold cross-validation. Finally, the simulation study also shows that the STR-Tree model is a promising alternative when out-of-sample prediction of unknown nonlinear functions is considered. When put into proof with real datasets, the STR-Tree model has a superior predictive ability than CART. Model averaging is discussed and the main result is that the combination of the STR-Tree model with the multivariate adaptive regression splines (MARS) of Friedman (1991) is a viable alternative to nonlinear prediction. A Matlab code for carrying out the modeling cycle exists and can be obtained from the authors.

The paper is divided as follows. In Section 2, we briefly introduce some important regression tree concepts and introduce the main notation. Section 3 describes the proposed model. Section 4 discusses the model building strategy and parameter estimation. The use of categorical data is considered in Section 5. A Monte Carlo Experiment is conducted in Section 6. Examples with six datasets are presented in Section 7. Finally, Section 8 concludes. A technical appendix provides the proofs of the theorems.

2. Regression Trees

A regression tree is a nonparametric model which looks for the best local prediction of a continuous response through the recursive partitioning of the space of the predictor variables. Usually, regression trees are estimated by a greedy recursive partitioning algorithm; see Breiman, Friedman, Olshen, and Stone (1984). The fitted model is displayed in a graph which has the format of a binary decision tree with *parent* and *terminal nodes* (also called *leaves*), and which grows from the *root node* to the terminal nodes. For example, Figure 1 displays a tree with three parent nodes and four leaves.

2.1. **Mathematical Formulation.** Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{mt})' \in \mathbb{X} \subseteq \mathbb{R}^m$ be a vector which contains m explanatory variables for a continuous univariate response $y_t \in \mathbb{R}$. The relationship between y_t and $\mathbf{x}_t$ follows the regression model

$$y_t = f(\mathbf{x}_t) + \varepsilon_t, \tag{1}$$

where the functional form $f(\cdot)$ is unknown and there are no assumptions about the distribution of the random term ε_t. Following Lewis and Stevens (1991), a regression tree model with K leaves is a recursive partitioning model that approximates $f(\cdot)$ by a general nonlinear function $H(\mathbf{x}_t; \boldsymbol{\psi})$ of $\mathbf{x}_t$ indexed by the vector of parameters $\boldsymbol{\psi} \in \mathbb{R}^r$; r is the total number of parameters. Frequently, $H(\cdot)$ is a piecewise constant function defined by K subregions $k_i(\boldsymbol{\theta}_i)$, $i = 1, \ldots, K$, of some domain $\mathbb{K} \subset \mathbb{R}^m$. Each region is determined by the parameter vector $\boldsymbol{\theta}_i$, $i = 1, \ldots, K$, such that

$$f(\mathbf{x}_t) \approx \sum_{i=1}^{K} \beta_i I_i(\mathbf{x}_t; \boldsymbol{\theta}_i), \tag{2}$$

where

$$I_i(\mathbf{x}_t; \boldsymbol{\theta}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_t \in k_i(\boldsymbol{\theta}_i); \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and $\boldsymbol{\psi} = (\beta_1, \ldots, \beta_K, \boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K)'$. Conditionally to the knowledge of the subregions, the relationship between y_t and $\mathbf{x}_t$ in (1) is approximated by a linear regression on a set of K dummy variables.

The most important reference in regression tree models is the CART approach discussed in Breiman, Friedman, Olshen, and Stone (1984). In this context, it is usual to define the subregions k_i, $i = 1, \ldots, K$, in (2) by hyperplanes that are orthogonal to the axis of the predictor variables. For example, consider a simple tree structure with $K = 2$ leaves and

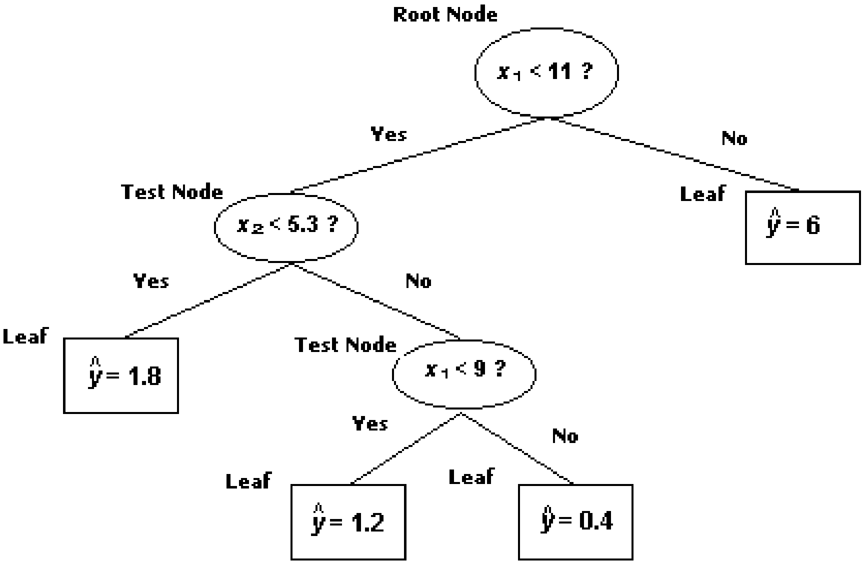


FIGURE 1. Graphical display of a regression tree.

depth $d = 1$. The unknown function $f(\mathbf{x}_t)$ in (1) may be approximated by a constant model in each leaf, written as

$$y_t = \beta_1 I(\mathbf{x}_t; s_0, c_0) + \beta_2 \left[1 - I(\mathbf{x}_t; s_0, c_0)\right] + \varepsilon_t, \tag{4}$$

where

$$I(\mathbf{x}_t; s_0, c_0) = \begin{cases} 1 & \text{if } x_{s_0 t} \le c_0; \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

$s_0 \in \mathbb{S} = \{1, 2, \dots, m\}$, and $x_{s_0 t} \in \mathbf{x}_t$.

To mathematically represent more complex tree structures, we adopt a labeling scheme which is similar to the one used in Denison, Mallik, and Smith (1998). The root node is at position 0 and a parent node at position j generates the left-child node and right-child node at positions $2j + 1$ and $2j + 2$, respectively. Consider a tree with N parent nodes. The variables x_{s_j}, $j = 1, \dots, N$ are usually called *splitting variables*.

3. Tree-Structured Smooth Transition Regression (STR-Tree)

The main idea of the STR-Tree model is to take advantage of the CART structure, but also to introduce elements which make it feasible to use standard inferential procedures. Whenever possible, we intend to keep the interpretability of the tree-based models. The highly discontinuous functional form of the model fitted by the CART and the strategy to decrease the sum of squared errors by splitting the sample recursively, pose a problem to test the significance of the model and to make classical inference. The idea here is the same used in Suárez and Lutsko (1989): the substitution of sharp splits in the CART model by smooth splits. Consider the simplest tree with two terminal nodes generated as in (4). If we replace the indicator function $I(\cdot)$ in (4) by a logistic function defined as

$$G(\mathbf{x}_t; s_0, \gamma_0, c_0) = \frac{1}{1 + e^{-\gamma_0\left(x_{s_0 t} - c_0\right)}}, \tag{6}$$

we obtain $y_t = \beta_1 G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \beta_2 \left[1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)\right] + \varepsilon_t$, where now we have the additional parameter γ_0, called the *slope parameter*, which controls the smoothness of the logistic function. This change causes an important difference from the CART approach: splitting the root node will not separate two subsets of observations but it will create two fuzzy sets (Zadeh 1965) where all observations will belong to, but with a different degree of membership. When the slope parameter approaches zero, it leads to the fuzziest situation in which there is no gain in splitting the data. The parameter c_0 is called the *location parameter*. When the transition is very smooth the model looses the standard tree interpretability. However, the STR-Tree model can be seen as a fuzzy regression model or a model where we associate a probability of being in each regime.

As the CART node partition is nested in the smooth transition approach as a special case when the slope parameter approaches infinity, we argue that the STR-Tree model inherits all the function approximation properties of the regression-trees with sharp splits.

Replacing the sharp splits by smooth ones has some advantages. Firstly, standard inferential theory can be used to test hypothesis about the location of the splits and to construct confidence intervals to the predictions. Moreover, as shown in the simulations in Section 6, cross-validation is inefficient in specifying correct-sized trees and is computationally intensive. Finally, smoothing between adjacent nodes can reduce bias and variance in the predictions, specially near the node boundaries. Finally, it is possible to interpret the regression tree approach as a particular case of the STR models discussed in Chan and Tong (1986) and Granger and Teräsvirta (1993).

4. Model Building

The architecture of tree-based models is usually determined from the data. Popular methods for doing that are based of cross-validation or information criteria. Applying an information criterion (IC) to decide whether or not another a given node should be split or not requires estimation of a more complex model (with one more split). In this situation the larger model is not identified and its parameters cannot be estimated consistently. This is likely to cause numerical problems in maximum likelihood estimation. Besides, even when convergence is achieved, lack of identification causes a severe problem in interpreting the IC. The tree model with more terminal nodes (splits) is nested in the model with less terminal nodes. A typical IC comparison of the two models is then equivalent to a likelihood ratio test, see, for example, Teräsvirta and Mellin (1986) for discussion. The choice of the IC determines the (asymptotic) significance level of the test. But then, when the larger model is not identified under the null hypothesis, the likelihood ratio statistic does not have its customary asymptotic χ^2 distribution when the null holds. For more discussion of the general situation of a model only being identified under the alternative hypothesis, see, for example, Davies (1977, 1987) and Hansen (1996).

Here we adopt a different strategy following the modeling cycle described in Teräsvirta (1994), Medeiros and Veiga (2005), and Medeiros, Teräsvirta, and Rech (2006). The "architecture" of the model has to be determined from the data and we call this stage *specification* of the model, which involves two decisions: the selection of the node to be split and the index of the splitting variable. The specification stage will be carried out by a sequence of Lagrange Multiplier (LM) tests following the ideas originally presented in Luukkonen, Saikkonen, and Teräsvirta (1988). An alternative approach based on 10-fold cross-validation is also possible; however the computational burden involved is dramatically high. The specification stage also requires *estimation* of the parameters of the model. What follows thereafter is *evaluation* of the final estimated model. Tree models are usually evaluated by their out-of-sample performance (predictive ability). In this paper we follow the literature and evaluate the STR-Tree model in the same way. The construction of misspecification tests for the STR-Tree model in the same spirit of Eitrheim and Teräsvirta (1996) is also possible, but this topic is beyond the scope of the paper.

Following the "specific-to-general" principle, we start the cycle from the root node (depth 0) and the general steps are:

(1) Specification of the model by selecting in the depth d, using the LM test, a node to be split (if not in the root node) and a splitting variable.

(2) Parameter estimation.

(3) Evaluation of the estimated model by checking if it is necessary to: (a) Change the node to be split; (b) change the splitting variable; and (c) remove the split.

(4) Use the final tree model for prediction or descriptive purposes.

The modeling cycle begins from the root node (depth 0) by testing the null hypothesis of a global constant model against a STR-Tree model with only 2 terminal nodes.

4.1. **Parameter Estimation.** Consider a full-grown STR-Tree model with depth d, $K = 2^d$ terminal nodes (leaves), and $N = \sum_{i=1}^{d} 2^i$ parent nodes, defined as

$$y_t = H(\mathrm{x}_t;\boldsymbol{\psi}) + \varepsilon_t = \sum_{k=1}^{K} \beta_{K+k-2} B_k(\mathrm{x}_t;\boldsymbol{\theta}_k) + \varepsilon_t, \tag{7}$$

where $H(\mathrm{x}_t;\boldsymbol{\psi}) = \sum_{k=1}^{K} \beta_{K+k-2} B_k(\mathrm{x}_t;\boldsymbol{\theta}_k)$ and $B_k(\mathrm{x}_t;\boldsymbol{\theta}_k)$, $k = 1,\ldots,K$, is defined by products of the logistic function. The parameter vector $\boldsymbol{\psi} = (\beta_{K-1},\ldots,\beta_{2K-2},\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_K')'$ has $r = K + 2N$ elements. As an example, consider a STR-Tree model with depth $d = 2$,

$K = 4$, $N = 3$, and functions $B_k(\mathrm{x}_t; \boldsymbol{\theta}_k)$, $k = 1, \ldots, K$, in (7) written as

$$B_1\left(\mathrm{x}_t; \boldsymbol{\theta}_1\right) = G\left(\mathrm{x}_t; s_0, \gamma_0, c_0\right) G\left(\mathrm{x}_t; s_1, \gamma_1, c_1\right);$$

$$B_2\left(\mathrm{x}_t; \boldsymbol{\theta}_2\right) = G\left(\mathrm{x}_t; s_0, \gamma_0, c_0\right) \left[1 - G\left(\mathrm{x}_t; s_1, \gamma_1, c_1\right)\right];$$

$$B_3\left(\mathrm{x}_t; \boldsymbol{\theta}_3\right) = \left[1 - G\left(\mathrm{x}_t; s_0, \gamma_0, c_0\right)\right] G\left(\mathrm{x}_t; s_2, \gamma_2, c_2\right); \text{ and}$$

$$B_4\left(\mathrm{x}_t; \boldsymbol{\theta}_4\right) = \left[1 - G\left(\mathrm{x}_t; s_0, \gamma_0, c_0\right)\right] \left[1 - G\left(\mathrm{x}_t; s_2, \gamma_2, c_2\right)\right].$$

The total number of parameters to be estimated is 10 and there are three splitting variables to be selected. It is important to stress that all tree architectures can be seen as a restricted version of a full grown tree, which is used here just to make the presentation clearer.

4.1.1. *Main Assumptions.* At this point we have to make the following set of assumptions.

ASSUMPTION 1. *The sequence $\{\mathrm{x}_t\}_{t=1}^{T}$ is formed by independent and identically distributed (IID) random vectors and have a common joint distribution $\mathcal{D}$ on $\boldsymbol{\Delta}$, a measurable Euclidean space, with measurable Radon-Nikodým density.*

ASSUMPTION 2. *The sequence $\{\varepsilon_t\}_{t=1}^{T}$ is formed by independent and normally distributed (NID) random variables with zero mean and variance $\sigma^2 < \infty$, that is $\varepsilon_t \sim$ NID $(0, \sigma^2)$.*

ASSUMPTION 3. *The $r \times 1$ true parameter vector $\boldsymbol{\psi}^*$ is an interior point of the compact parameter space $\boldsymbol{\Psi}$ which is a subspace of $\mathbb{R}^r$, the r-dimensional Euclidean space.*

ASSUMPTION 4. *The parameters $\gamma_i > 0$, $i = 1, \ldots, N$, where N is the number of parent nodes. Furthermore, if for two adjacent parent nodes at positions $2j + 1$ and $2j + 2$, $x_{s_{2j+1}t} = x_{s_{2j+2}t}$, then $c_{s_{2j+1}} < c_{s_{2j+2}}$.*

Assumption 1 states that we are working with IID data such as cross-sectional or a set of time-series with IID observations. Although Assumption 2 may seem a little restrictive, model (7) is still very flexible. Furthermore, Assumption 2 allows us to work in a maximum likelihood framework that will be equivalent to nonlinear least-squares. In the case of non-Gaussian errors, Assumption 2 may be substituted by some moment conditions and a quasi-maximum likelihood framework should be used instead. The main difference will be related to the computation of the covariance matrix of the parameter estimates. In addition, a robust version of the tests presented latter can be constructed in the same spirit of Wooldridge (1991) and Medeiros, Teräsvirta, and Rech (2006). Assumption 3 is standard and Assumption 4 guarantees that the STR-Tree model is identifiable.

As discussed previously, we estimate the parameters of our STR-Tree model by maximum likelihood (ML). The use of maximum likelihood makes it possible to obtain an idea of the uncertainty in the parameter estimates through (asymptotic) standard deviation estimates. The STR-Tree model is similar to many linear or nonlinear models in that the information matrix of the log-likelihood function is block diagonal in such a way that we can concentrate the likelihood and first estimate the parameters of the conditional mean. Conditional maximum likelihood is thus equivalent to nonlinear least squares (NLS).

The nonlinear least squares estimator (NLSE) of the parameters equals

$$\widehat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \frac{1}{T} Q_T(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T} q_t(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2. \tag{8}$$

4.1.2. *Existence.* The proof of existence of the NLSE is based on Lemma 2 of Jennrich (1969), which establishes that under certain conditions of continuity and measurability on the mean square error (MSE) function, the NLSE as in (8) exists. Theorem 1 states the necessary conditions for the existence of the NLSE.

THEOREM 1. *The STR-Tree model satisfies the following conditions and the NLSE exists.*

(1) *For each* $\mathbf{x}_t \in \mathbb{X} \subseteq \mathbb{R}^m$, *function* $H_{\mathbf{x}}(\boldsymbol{\psi}) = H(\mathbf{x}_t; \boldsymbol{\psi})$ *is continuous in compact subset* Ψ *of the Euclidean space.*

(2) *For each* $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^r$, *function* $H_{\boldsymbol{\psi}}(\mathbb{X}) = H(\mathbf{x}_t; \boldsymbol{\psi})$ *is measurable in space* $\mathbb{X}$.

(3) $\varepsilon_t \sim$ *IID*$(0, \sigma^2)$.

4.1.3. *Consistency.* The consistency of the NLSE was proved in Jennrich (1969). We follow Amemiya (1983) and state the following theorem.

THEOREM 2. *Under the Assumptions 1–5 $\widehat{\boldsymbol{\psi}}$ is strong consistent for $\boldsymbol{\psi}^*$, i.e., $\widehat{\boldsymbol{\psi}} \overset{a.s.}{\rightarrow} \boldsymbol{\psi}^*$.*

4.1.4. *Asymptotic Normality.* Asymptotically normality of the NLSE was also carefully proved in Jennrich (1969). We follow his results and the developments in Amemiya (1983) and state the following theorem.

THEOREM 3. *Under the Assumptions 1–5*

$$T^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \overset{d}{\rightarrow} N\left(\mathbf{0}, -\underset{T\rightarrow\infty}{plim}\, \mathbf{A}(\boldsymbol{\psi}^*)^{-1}\right), \tag{9}$$

where $\mathbf{A}(\boldsymbol{\psi}^*) = \frac{1}{\sigma^2}\frac{\partial^2 Q_T(\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}$.

REMARK 1. *The extension of the above theorems to the case of non-IID observations and to misspecified models is relatively straightforward. The results of White (1982), White (1994), and Wooldridge (1994) can be applied.*

4.1.5. *Concentrated Least-Squares.* Conditional on the knowledge of the parameters $\boldsymbol{\theta}_k$ in (7), $k = 1, \ldots, K$, model (7) is just a linear regression and the vector of parameters $\boldsymbol{\beta} = (\beta_{K-1}, \ldots, \beta_{2K-2})'$ can be estimated by ordinary least-squares (OLS) as

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{B}(\boldsymbol{\theta})'\mathbf{B}(\boldsymbol{\theta})\right]^{-1}\mathbf{B}(\boldsymbol{\theta})'\mathbf{y}, \tag{10}$$

where $\mathbf{y} = (y_1, \ldots, y_T)'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$, and

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{pmatrix} B_1(\mathbf{x}_1;\boldsymbol{\theta}_1) & \cdots & B_K(\mathbf{x}_1;\boldsymbol{\theta}_K) \\ \vdots & \ddots & \vdots \\ B_1(\mathbf{x}_T;\boldsymbol{\theta}_1) & \cdots & B_K(\mathbf{x}_T;\boldsymbol{\theta}_K) \end{pmatrix}.$$

The parameters $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$, are estimated conditionally on $\boldsymbol{\beta}$ by applying the Levenberg-Marquadt algorithm which completes the ith iteration.

4.2. **Splitting the Nodes.** We have a particular interest in the hypothesis concerning the significance of splitting the root node. If we re-parameterize the STR-Tree model as:

$$y_t = \phi_0 + \lambda_0 G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \varepsilon_t, \tag{11}$$

where $\phi_0 = \beta_2$ and $\lambda_0 = \beta_1 - \beta_2$, we obtain a more parsimonious representation of the simplest STR-Tree model. In order to test the significance of the first split, a convenient null hypothesis is $\mathcal{H}_0 \, : \, \gamma_0 = 0$ against the alternative $\mathcal{H}_a \, : \, \gamma_0 > 0$. An equivalent null hypothesis is $\mathcal{H}'_0 \, : \, \lambda_0 = 0$. However, under $\mathcal{H}_0$, the nuisance parameters λ_0 and c_0 can assume different values without changing the likelihood function. This poses an identification problem whose solution was first discussed by Davies (1977).

We adopt as a solution for this problem the one proposed in Luukkonen, Saikkonen, and Teräsvirta (1988), that is to approximate the function $G(\cdot)$ by a third-order Taylor expansion around $\gamma = 0$. After some algebra we get

$$y_t = \alpha_0 + \alpha_1 x_{s_0,t} + \alpha_2 x^2_{s_0,t} + \alpha_3 x^3_{s_0,t} + e_t, \tag{12}$$

where α_i, $i = 0, 1, 2, 3$, is a parameter that is function of γ_0, c_0, ϕ_0, and λ_0, $e_t = \varepsilon_t + \lambda_0 R(\mathbf{x}_t; s_0, \gamma_0, c_0)$, and $R(\mathbf{x}_t; s_0, \gamma_0, c_0)$ is the remainder. Thus,

$$\mathcal{H}_0 : \, \alpha_i = 0, \; i = 1, 2, 3. \tag{13}$$

Note that under $\mathcal{H}_0$, the remainder of the Taylor expansion vanishes and $e_t = \varepsilon_t$, so that the properties of the error process remain unchanged under the null and thus asymptotic inference can be used. Finally, one may also view (12) as resulting from a local approximation to the log-likelihood function, which for observation t takes the form

$$l_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{y_t - \alpha_0 - \alpha_1 x_{s_0,t} - \alpha_2 x^2_{s_0,t} - \alpha_3 x^3_{s_0,t}\right\}^2. \tag{14}$$

At this point we make the following additional assumption.

ASSUMPTION 5. *$E|x_{s_0t}|^{\delta} < \infty$, $\forall\ s_0 \in \mathbb{S}$, for some $\delta > 6$.*

This enables us to state the following well-known result.

THEOREM 4. *Under $\mathcal{H}_0 : \gamma_0 = 0$ and Assumptions (2)–(5), the LM type statistic*

$$LM = \frac{1}{\widehat{\sigma}^2} \sum_{t=1}^{T} \widehat{\varepsilon}_t \boldsymbol{\nu}_t' \left\{ \sum_{t=1}^{T} \boldsymbol{\nu}_t \boldsymbol{\nu}_t' - \sum_{t=1}^{T} \boldsymbol{\nu}_t \mathbf{h}_t' \left(\sum_{t=1}^{T} \mathbf{h}_t \mathbf{h}_t' \right)^{-1} \sum_{t=1}^{T} \mathbf{h}_t \boldsymbol{\nu}_t' \right\}^{-1} \sum_{t=1}^{T} \boldsymbol{\nu}_t \widehat{\varepsilon}_t, \quad (15)$$

where $\widehat{\varepsilon}_t = y_t - \widehat{\beta}_0$ is the estimated residuals under the null, $\widehat{\sigma}^2 = (1/T)\sum_{t=1}^{T} \widehat{\varepsilon_t}^2$, $\mathbf{h}_t = 1$, and $\boldsymbol{\nu}_t = \left(x_{s_0t}, x_{s_0t}^2, x_{s_0t}^3\right)'$, has an asymptotic χ^2 distribution with 3 degrees of freedom.

REMARK 2. *Note that, under $\mathcal{H}_0$, $\widehat{\beta}_0 = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} E\left(y_t\right)$.*

Until this point, we have just interpreted the simplest tree model as a particular case of the STR model as in Granger and Teräsvirta (1993) and the testing strategy to split the root node corresponds to a linearity test in which the linear model in question is a global constant model. However, the key idea is to consider the basic testing procedure described above in a more complex framework. To give an example of a more complex model, consider that the null hypothesis (13) was rejected and a STR-Tree model with two leaves was consistently estimated. A natural way, within the tree framework, of considering a hypothesis of misspecification is by formulating a new model that splits one between the two created nodes, say the left child node, leading to the following model

$$\begin{aligned} y_t =& H(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t \\ =& \left\{\beta_3 G(\mathbf{x}_t; s_1, \gamma_1, c_1) + \beta_4 \left[1 - G(\mathbf{x}_t; s_1, \gamma_1, c_1)\right]\right\} G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \quad (16) \\ & \beta_2 \left[1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)\right] + \varepsilon_t. \end{aligned}$$

Therefore, rewriting (16) as

$$y_t = \left[\phi_1 + \lambda_1 G(\mathbf{x}_t; s_1, \gamma_1, c_1)\right] G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \beta_2 \left[1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)\right] + \varepsilon_t, \quad (17)$$

where $\phi_1 = \beta_3$ and $\lambda_1 = \beta_3 - \beta_4$, a convenient null hypothesis is $\mathcal{H}_0 :\ \gamma_1 = 0$.

However, under the null hypothesis, the model (17) can not be consistently estimated because of the nuisance parameters λ_1 and c_1. For solving this identification problem, we proceed as before and approximate the function $G(\cdot)$ by its third-order Taylor expansion around $\mathcal{H}_0$. After some algebra we get

$$\begin{aligned} y_t = &\alpha_0 + \alpha_1 G\left(x_{s_0 t};\gamma_0, c_0\right) + \alpha_2 G\left(x_{s_0 t};\gamma_0, c_0\right) x_{s_1 t} + \\ &\alpha_3 G\left(x_{s_0 t};\gamma_0, c_0\right) x_{s_1 t}^2 + \alpha_4 G\left(x_{s_0 t};\gamma_0, c_0\right) x_{s_1 t}^3 + e_t, \end{aligned} \tag{18}$$

where $e_t = \varepsilon_t + R\left(\mathbf{x}_t; s_1, \gamma_1, c_1\right)$; $R\left(\mathbf{x}_t; s_1, \gamma_1, c_1\right)$ is the remainder. The decision for splitting the node corresponds to the rejection of the following null hypothesis

$$\mathcal{H}_0 : \alpha_i = 0, \; i = 2, 3, 4. \tag{19}$$

The test statistic is (15) with

$$\mathbf{h}_t = \left(1,\, G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right),\, \widehat{\alpha}_1 \left.\frac{\partial G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right)}{\partial \gamma_0}\right|_{\mathcal{H}_0},\, \widehat{\alpha}_1 \left.\frac{\partial G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right)}{\partial c_0}\right|_{\mathcal{H}_0}\right)'$$

and $\boldsymbol{\nu}_t = \left(G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right) x_{s_1 t},\, G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right) x_{s_1 t}^2,\, G\left(x_{s_0 t};\widehat{\gamma}_0, \widehat{c}_0\right) x_{s_1 t}^3\right)'$.

From the assumption of normality of the error term, the information matrix is block diagonal and thus we can assume that the error variance is fixed. The test can be carried out according to the following steps:

(1) Estimate the STR-Tree model under the null hypothesis $\mathcal{H}_0$ and compute the residuals $\widehat{\varepsilon}_t$. Compute the sum of the squared residuals $SSR_0 = \sum_{t=1}^{T} \widehat{\varepsilon}_t^2$.

(2) Regress $\widehat{\varepsilon}_t$ on $\mathbf{h}_t$ and $\boldsymbol{\nu}_t$. Compute the sum of squared residuals obtained from this regression (SSR_1).

(3) Compute the χ^2 statistic

$$LM_\chi = T\frac{SSR_0 - SSR_1}{SSR_0}, \tag{20}$$

or the F version of the test

$$LM_F = \frac{(SSR_0 - SSR_1)/3}{SSR_1/(T-7)}, \tag{21}$$

where T is the sample size. Under the null LM_χ is asymptotically distributed as a χ^2 distribution with 3 degrees of freedom and LM_F has an asymptotic F distribution with 3 and $T-7$ degrees of freedom.

Hereafter, the idea is to carry out a sequence of LM-type tests to grow the tree model in the same format as the one presented above and the general form of the test statistic when testing a model with j nodes against an alternative with $j+1$ nodes is given by:

$$LM = \frac{(SSR_0 - SSR_1)/3}{SSR_1/[T-(p+3)]}, \tag{22}$$

where p is the total number of elements of the vector $\mathbf{h}_t$.

4.2.1. *Modeling Cycle from the root node (depth 0).* The decision to split the root node is based on the following steps.

(1) For each explanatory variable, apply the LM-type test described above and select the variable x_{s_0t} that generates the lowest p-value below a specified level α. In case of all candidate variables do not produce a significant split, the root node is declared as terminal and the global constant model is selected as the best model. Otherwise, two children nodes are generated to compose the first depth of the tree.
(2) Conditional to the choice of s_0, estimate the vector of parameters $\boldsymbol{\psi} = (\gamma_0, c_0, \beta_1, \beta_2)'$ by concentrated least squares.

4.2.2. *Modeling Cycle from the 1st depth.* After the tree has started to grow from the root node, the first depth is created and the cycle continues by testing for the adequacy of splitting one between the two children nodes. The null hypothesis in this test concerns the conditional linear model and the alternative brings the inclusion of a nonlinear term that is responsible for splitting the node. From now on, besides selecting a splitting variable, we shall also select which one between the two created nodes shall be split at the first place.

(1) For each combination of splitting variable index in $\mathbb{S} = \{1, 2, \ldots, m\}$ and node number in $\mathbb{D}_1 = \{1, 2\}$, apply the LM-type test and select the indexes $j_1 \in \mathbb{D}_1$ and $s_{j_1} \in \mathbb{S}$ that generates the lowest p-value below a pre-specified significance level. If there is no significant split, the tree growing process stops.

(2) Estimate the parameters of the model.

4.2.3. *Modeling Cycle from the kth depth.* The execution of the algorithm in a general depth k is straightforward.

(1) Apply the LM test to all combinations of splitting variables indexes and nodes in the set $\mathbb{D}_k$ which contains all numbers of children nodes that compose the kth depth. Note that $\mathbb{D}_k \subseteq \left\{2^k - 1, 2^k, \ldots, 2^{k+1} - 2\right\}$.

(2) Select $j_1 \in \mathbb{D}_k$ and $s_{j_k} \in \mathbb{S}$ by the rank of significant p-values obtained through the LM-type test.

(3) Estimate the parameters of the model.

The whole modeling cycle ends when a determined depth do not produce children nodes.

4.3. **Sequential Tests.** To achieve the final tree model, we perform a sequence of n correlated LM-type tests of hypothesis in which n is a random variable. Due to multiplicity from repeated significance testing, we have to control the overall type I error under the risk of an overstatement of the significance of the results (more splits are reported to be significant than it should be). To remedy this situation, we adopt the following procedure. For the nth test in the sequence, if it is performed in the dth depth the significance level is $\alpha(d, n) = \frac{\alpha}{n^d}$. In the root node $(d = 0)$ and we apply the first test $(n = 1)$ for splitting the node at a significance level α, if the null is rejected than the second $(n = 2)$ test is applied in the 1st depth $(d = 1)$ and the significance level is $\alpha/2$. By forcing the test to be more rigorous in deeper depths, we create a procedure that diminishes the importance of using post-pruning techniques.

There are several alternatives to control the overall size of the sequence of tests: Hochberg (1988), Benjamini and Hochberg (1995,1997), Benjamini and Yekutieli (2000,2001), and Benjamini and Liu (1999). However, by our experiments, our simple methodology works well and the comparison between different techniques to reduce the nominal size of each test is beyond the scope of the paper.

5. CATEGORICAL DATA

In principle, the previous developments do not take into account the case where some of the variables are categorical. However, the extension to include categorical data is straightforward. The main idea is to replace the constant model in each terminal node by a linear regression on a constant and a set of dummy variables representing the categorical data.

Let $\mathbf{x}_t = (\mathbf{z}_t', \mathbf{w}_t')'$, were $\mathbf{z}_t$ is a vector of categorical variables and $\mathbf{w}_t$ is a vector of continuous variables. Let $\mathbf{D}_t(\mathbf{z}_t)$ be a vector of dummy variables representing the categorical vector $\mathbf{z}_t$. In that case model (7) may be rewritten as:

$$y_t = H\left(\mathbf{x}_t; \boldsymbol{\psi}\right) + \varepsilon_t = \sum_{k=1}^{K} \boldsymbol{\beta}'_{K+i-1} \mathbf{D}_t(\mathbf{z}_t) B_k(\mathbf{w}_t; \boldsymbol{\theta}_k) + \varepsilon_t. \qquad (23)$$

6. MONTE CARLO EXPERIMENT

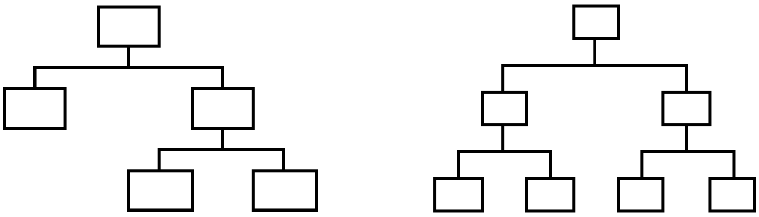
In this section study the small sample properties of the nonlinear least squares estimators under correct specification of the STR-Tree model and investigate the performance of three different tree-growing algorithms:

CART: We use the most traditional CART tree growing strategy. This consists of growing the tree using as a stopping rule the minimum of five observations per terminal node, and then prune the tree using the 1-SE rule with errors estimates obtained by 10-fold cross validation.

STR-Tree/LM: This strategy uses the LM test to select the node and splitting variable. This specification strategy does not need pruning and the control of the overall error is done by the reducing the test size during the tree growing.

STR-Tree/CV: We carry at each node a 10-fold cross-validation experiment to select the splitting variable that minimizes the overall MSE (Mean Square of Errors) evaluated out-of-sample.

We simulate two tree architectures which are illustrated in Figure 2, with different combinations of smoothness parameters. Thus, five models are simulated for Architecture I which contains three terminal nodes and three models are simulated for Architecture II which has four terminal nodes. Basically, we consider three types of splits (see Table 1): very smooth ($\gamma_i = 0.5$), moderate sharp ($\gamma_i = 5$), and sharp ($\gamma_i = \infty$). The sharp splits are used to evaluate the robustness of the STR-Tree model when the true specification is a regression-tree with hard thresholds. We also mix types of splits. Model 1.1, for example, is obtained from two consecutive smooth splits and Model 1.4 brings a smooth split at the root node, followed by a moderate sharp split.



(a) Architecture I (b) Architecture II

FIGURE 2. Small simulated trees architectures

We simulate 1000 replications for each model with sample sizes $T = 150$ and $T = 500$. As the main concern is about the effects of the slope parameter, there is not much variation in the choice of the constants within the nodes. Three uncorrelated and normally distributed predictor variables are used as candidates to be the splitting variables: $x_1 \sim \mathrm{N}(10, 2.56)$; $x_2 \sim \mathrm{N}(90, 9)$; and $x_3 \sim \mathrm{N}(25, 4)$. The error term is defined as $\varepsilon_t \sim \mathrm{N}(0, 1)$. Since the slope parameter is not scale-free, we standardize the argument of the logistic function,

TABLE 1. Smoothness of the splits in the STR-Tree simulations

	Model	First Split	Second Split	Third Split
Architecture I (3 leaves)	1.1	$\gamma_0 = 0.5$	$\gamma_2 = 0.5$	—
	1.2	$\gamma_0 = 5$	$\gamma_2 = 5$	—
	1.3	$\gamma_0 = 5$	$\gamma_2 = 0.5$	—
	1.4	$\gamma_0 = 0.5$	$\gamma_2 = 5$	—
	1.5	$\gamma_0 = \infty$	$\gamma_2 = \infty$	—
Architecture II (4 leaves)	2.1	$\gamma_0 = 0.5$	$\gamma_1 = 0.5$	$\gamma_2 = 0.5$
	2.2	$\gamma_0 = 5$	$\gamma_1 = 5$	$\gamma_2 = 5$
	2.3	$\gamma_0 = \infty$	$\gamma_1 = \infty$	$\gamma_2 = \infty$

dividing it by the standard deviation of the splitting variable. The other parameters are fixed according to Table 2. In the simulations concerning parameter estimation we do not consider the cases where $\gamma_i = \infty$.

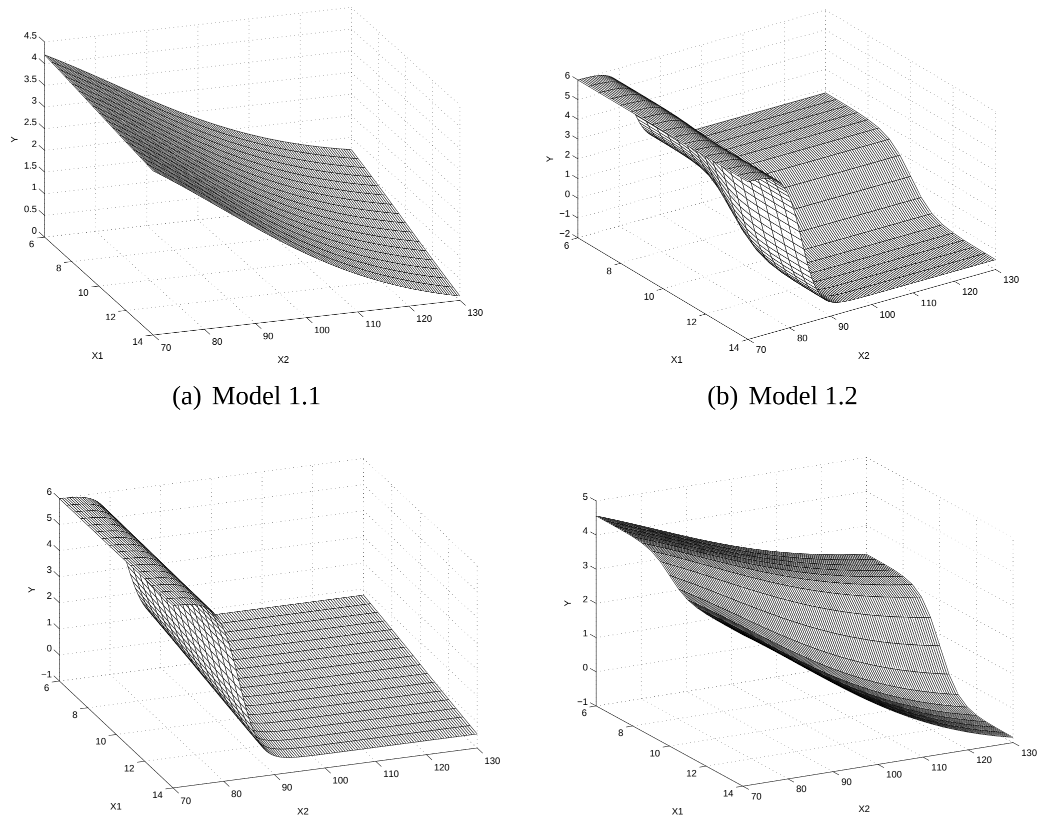
TABLE 2. Parameters in the simulated STR-Tree models

	Architecture I	Architecture II
Constants in the nodes	$\beta_1 = 6$ $\beta_5 = 1.8$; $\beta_6 = -1.5$	$\beta_3 = 6$; $\beta_4 = 3.2$ $\beta_5 = 1.8$; $\beta_6 = -1.5$
Location parameters	$c_0 = 83$; $c_2 = 10$	$c_0 = 90$; $c_1 = 10$; $c_2 = 25$
Indexes of splitting variables	$s_0 = 2$; $s_2 = 1$	$s_0 = 2$; $s_1 = 1$; $s_2 = 3$

As shown in Table 2, the location parameters are chosen strategically at median points for simulations under Architecture II. The aim is to provide a maximum amount of information within the created nodes. The only concern related to the choice of the constants within the nodes is to yield different local models.

The difference among models for Architecture I can be seen in Figure 3, which shows the response surface for each one of the simulated trees. When all splits are moderate sharp such as in model 1.2, the surface looks like a bivariate histogram. On the other hand, a sequence of extremely smooth splits (Model 1.1) produces a relationship between the response and regressors that is almost linear.

6.1. **Parameter Estimation.** In this Section, we discuss the empirical results obtained with the use of the NLSE in the simulated models. The results are described through descriptive statistics such as the sample mean. Two measures are chosen to evaluate the



(a) Model 1.1

(b) Model 1.2

(c) Model 1.3

(d) Model 1.4

FIGURE 3. Geometric Features of the Simulated Models (Architecture I)

variability of the estimates; the sample standard deviation and, as a more robust alternative, the median absolute deviation around the median (MAD):

$$MAD(\widehat{\psi}) = \text{median}\left(\left|\widehat{\psi} - \text{median}(\widehat{\psi})\right|\right). \tag{24}$$

Estimation of the slope parameter γ results in outliers and extreme values for some replications, hence the sample mean of the estimates is strongly affected by them. It is clear in Tables 3 and 4 that the parameter γ is strongly overestimated when $T = 150$. In these cases, the median seems to be a more robust measure of central tendency. Such problem does not occur with the location parameter, whose sample mean and median are close to the true value. Nevertheless, the variability of the location parameter estimator increases whenever there is a smooth split. As a consequence, the estimates of the parameters within the nodes are also affected, mainly in small samples. Thus, as it happened with Model 1.3, the sample mean and median for the local model estimates deviate from the population values. In general, the estimates, except for the smoothness parameter, are more precise in trees simulated with sharp splits. When mixing different types of splits, the results pointed out that a smooth split followed by a sharp split produces better results. In this situation, there are more observations left to be modeled after the first split. Finally, an important aspect of the simulation study is the indication that the NLS estimates converged, as expected, to the true value of the parameter whenever the sample size increases.

6.2. **Tree Architecture Specification by Different Algorithms.** We show in Tables 5 and 6, the performance of the three algorithms to identify the simulated STR-Tree models.

When all partitions involve only moderate sharp splits, the STR-Tree models yield more than 95% of correct specifications, independently of the simulated architecture. When $T = 150$, the sequence of LM tests produced significantly better results than 10-fold cross-validation. For $T = 500$ the performance of both are comparable, being the LM test slightly better. On the other hand, all strategies faced more trouble to specify correctly trees which were grown from very smooth splits. A very smooth split followed by a sharp one increased the number of misspecifications. However, the STR-Tree model specified by the LM test outperforms its competitors in most of the cases. The decision to generate trees with a highly smooth transition function at the first node turned the specification task very difficult for all algorithms, even so the STR-Tree/LM could perform quite satisfactorily in large samples. The main problem for this algorithm occurred in the situation involving a very smooth split at the root node followed by a sharp split in the subsequent node. It could specify neither the tree architecture nor the splitting variables. Whenever the CART algorithm is submitted to specify smooth trees, it tends to create less nodes than expected. In the opposite situation where the splits are moderate sharp, even the post-pruning procedure was not able to avoid overfitting. The strategy to use a 10-fold cross-validation experiment

TABLE 3. Descriptive Statistics for Estimation in Architecture I

Model 1.1	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (0.5)	0.518	0.112	0.502	0.066	0.503	0.055	0.498	0.036
$\hat{c}_0$ (83)	82.988	0.476	83.002	0.313	83.010	0.236	83.015	0.150
$\hat{\gamma}_2$ (0.5)	25183	207.942	0.570	0.268	0.533	0.178	0.522	0.113
$\hat{c}_2$ (10)	10.020	2.255	10.036	0.694	10.032	0.812	10.006	0.372
$\hat{\beta}_1$ (6)	6.016	0.364	6.007	0.229	6.004	0.173	5.996	0.113
$\hat{\beta}_5$ (1.8)	2.187	1.531	1.734	0.526	1.895	0.567	1.766	0.252
$\hat{\beta}_6$ (-1.5)	-1.915	1.510	-1.452	0.512	-1.623	0.630	-1.472	0.250
Model 1.2	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (5)	17.059	60.519	5.254	2.297	6.190	9.580	5.154	1.126
$\hat{c}_0$ (83)	83.035	0.183	83.019	0.097	83.008	0.071	83.002	0.042
$\hat{\gamma}_2$ (5)	35.672	319.697	5.581	1.642	11.260	153.725	5.158	0.767
$\hat{c}_2$ (10)	10.002	0.099	10.004	0.066	9.998	0.051	9.997	0.035
$\hat{\beta}_1$ (6)	6.012	0.189	6.013	0.128	5.996	0.106	5.998	0.072
$\hat{\beta}_5$ (1.8)	1.789	0.159	1.792	0.105	1.799	0.088	1.801	0.056
$\hat{\beta}_6$ (-1.5)	-1.501	0.161	-1.497	0.102	-1.501	0.087	-1.496	0.058
Model 1.3	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (5)	10.917	21.949	5.288	1.693	5.852	9.369	5.100	0.870
$\hat{c}_0$ (83)	83.006	0.146	82.998	0.073	82.999	0.061	82.998	0.040
$\hat{\gamma}_2$ (0.5)	16.131	126.012	0.542	0.238	0.526	0.171	0.520	0.107
$\hat{c}_2$ (10)	10.062	2.1281	9.969	0.707	10.003	0.964	10.007	0.368
$\hat{\beta}_1$ (6)	6.009	0.193	6.007	0.126	5.999	0.102	5.998	0.064
$\hat{\beta}_5$ (1.8)	2.204	1.420	1.766	0.509	1.953	0.739	1.785	0.243
$\hat{\beta}_6$ (-1.5)	-1.955	1.595	-1.441	0.464	-1.653	0.732	-1.483	0.246
Model 1.4	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (0.5)	0.527	0.145	0.505	0.0709	0.506	0.066	0.503	0.043
$\hat{c}_0$ (83)	83.045	0.513	83.023	0.342	83.011	0.277	83.020	0.183
$\hat{\gamma}_2$ (5)	45.670	386.809	5.402	1.779	9.213	110.411	5.077	0.741
$\hat{c}_2$ (10)	10.002	0.111	10.005	0.072	9.999	0.051	9.999	0.032
$\hat{\beta}_1$ (6)	6.004	0.357	5.984	0.223	6.000	0.188	5.994	0.123
$\hat{\beta}_5$ (1.8)	1.778	0.182	1.789	0.117	1.791	0.096	1.795	0.066
$\hat{\beta}_6$ (-1.5)	-1.511	0.219	-1.505	0.145	-1.503	0.116	-1.500	0.078

during the specification seems to produce results in the STR-Tree algorithm which are similar to CART ones. Although the overfitting is not so dramatic as in the CART case, when the splits are moderate sharp, the algorithm tended to create, mainly in small samples, trees which are larger than expected. With large samples and moderate sharp splits, the specification performance is comparable to the one done by the sequence of LM-type tests, but the computational burden is considerably high.

TABLE 4. Descriptive Statistics for Estimation in Architecture II

Model 2.1	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (0.5)	0.693	3.261	0.510	0.075	0.508	0.068	0.504	0.043
$\hat{c}_0$ (90)	90.017	0.542	89.998	0.380	89.999	0.305	89.994	0.198
$\hat{\gamma}_1$ (0.5)	46.594	397.130	0.805	0.531	17.417	253.162	0.557	0.188
$\hat{c}_1$ (10)	10.104	2.135	10.095	1.028	9.962	1.311	9.942	0.578
$\hat{\gamma}_2$ (0.5)	11.897	171.624	0.549	0.197	0.535	0.173	0.512	0.100
$\hat{c}_2$ (25)	24.997	1.953	25.031	0.781	24.990	0.710	24.997	0.358
$\hat{\beta}_3$ (6)	6.104	1.215	5.730	0.479	6.132	0.762	5.956	0.344
$\hat{\beta}_4$ (3.2)	3.045	1.261	3.429	0.445	3.102	0.729	3.271	0.323
$\hat{\beta}_5$ (1.8)	2.061	1.155	1.773	0.372	1.854	0.437	1.797	0.201
$\hat{\beta}_6$ (-1.5)	-1.777	1.091	-1.520	0.423	-1.555	0.451	-1.491	0.205
Model 2.2	$T = 150$				$T = 500$			
	Mean	Std. Dev.	Median	MAD	Mean	Std. Dev.	Median	MAD
$\hat{\gamma}_0$ (5)	70.192	1276.098	5.530	2.998	25.373	238.576	5.080	1.389
$\hat{c}_0$ (90)	90.009	0.238	90.002	0.130	90.003	0.116	89.997	0.065
$\hat{\gamma}_1$ (5)	104.765	527.811	6.993	3.932	367.216	4863.138	5.471	1.470
$\hat{c}_1$ (10)	9.997	0.157	9.999	0.094	10.005	0.082	10.006	0.056
$\hat{\gamma}_2$ (5)	76.126	553.641	6.700	3.596	55.747	323.296	5.207	1.261
$\hat{c}_2$ (25)	24.995	0.182	24.999	0.103	25.001	0.085	24.999	0.053
$\hat{\beta}_3$ (6)	6.004	0.218	6.004	0.132	5.990	0.124	5.988	0.084
$\hat{\beta}_4$ (3.2)	3.210	0.209	3.216	0.139	3.213	0.115	3.216	0.073
$\hat{\beta}_5$ (1.8)	1.790	0.194	1.782	0.125	1.789	0.099	1.794	0.067
$\hat{\beta}_6$ (-1.5)	-1.494	0.204	-1.487	0.128	-1.492	0.116	-1.493	0.079

TABLE 5. Percentage of Correct Specifications in Trees Simulated for Architecture I

	$T = 150$			$T = 500$		
Smoothness Parameters	CART	STR-Tree/LM	STR-Tree/CV	CART	STR-Tree/LM	STR-Tree/CV
γ_0=0.5 γ_2=0.5	7.7%	34.7%	6.4%	23%	84.2%	15.1%
γ_0=5 γ_2=5	8.4%	98.4%	89.9%	0%	97.8%	96.3%
γ_0=5 γ_2=0.5	16.4%	85.4%	42.5%	0.1%	99.1%	80.6%
γ_0=0.5 γ_2=5	37.8%	45.8%	38.4%	3.5%	6.1%	11.4%
$\gamma_0 = \infty$ $\gamma_2 = \infty$	92.2%	98.5%	18.7%	100%	99.5%	11.6%

TABLE 6. Percentage of Correct Specifications in Trees Simulated for Architecture II

	$T = 150$			$T = 500$		
Smoothness Parameters	CART	STR-Tree/LM	STR-Tree/CV	CART	STR-Tree/LM	STR-Tree/CV
$\gamma_0 = 0.5$ $\gamma_1 = 0.5$ $\gamma_2 = 0.5$	0.8%	4%	0.6%	4.3%	61.1%	1.3%
$\gamma_0 = 5$ $\gamma_1 = 5$ $\gamma_2 = 5$	25.9%	98.3%	76.7%	0%	98%	94.8%
$\gamma_0 = \infty$ $\gamma_1 = \infty$ $\gamma_2 = \infty$	96.3%	98.6%	16.0%	100%	99.4%	8.7%

Finally, when the true model is a regression-tree with hard splits, the CART algorithm, as expected, performs very well. However, the STR-tree model specified with the LM strategy is also very accurate, correctly selecting the true architecture in almost all replications. On the other hand, the 10-fold cross-validation is not a viable alternative to build STR-Tree models when the splits are hard. Surprisingly, when the splits change from moderate sharp ($\gamma_i = 5$) to sharp ($\gamma_i = \infty$), the performance of the CART algorithm improves dramatically.

6.3. **Out-of-Sample Predictions.** In order to evaluate the out-of-sample performance of the STR-Tree model we conduct the following experiment. We simulated 1000 replications with 750 observations of the following models:

- Model 1: Equation (66) in Friedman (1991)

$$y_i = \frac{40 \times \exp\left\{8\left[(x_{1i}-0.5)^2+(x_{2i}-0.5)^2\right]\right\}}{\exp\left\{8\left[(x_{1i}-0.2)^2+(x_{2i}-0.7)^2\right]\right\}+\exp\left\{8\left[(x_{1i}-0.7)^2+(x_{2i}-0.2)^2\right]\right\}} + \varepsilon_i,$$

  where ε_i is drawn from a standard normal distribution and x_{1i} and x_{2i} are drawn from a uniform distribution in the unit square.

- Model 2: Neural Network with three hidden units

$$\begin{aligned} y_i = 1.3 + 2.2f(1.5(0.5x_{1i} + 0.6x_{2i} - 10.5x_{3i} + 50)) - 1.7f(1.2(8.3x_{1i} + 0.2x_{3i} - 5)) \\ + 0.9f(5(0.7x_{1i} - 6.8x_{2i} + 3)) + \varepsilon_i, \end{aligned}$$

  where $f(z) = \left[1 + \exp\left(-z\right)\right]^{-1}$, x_{1i} and x_{2i} are drawn form a uniform distribution in the unit square, x_{3i} is drawn from a normal distribution with mean 5 and standard deviation 4, and ε_i is normally distributed with zero mean and unit variance.

- Model 3: Example 1 in Fan and Zhang (1999)

$$y_i = \sin(60x_{1i})x_{2i} + 4x_{1i}(1 - x_{1i})x_{3i} + \varepsilon_i,$$

  where x_{1i} is follows a uniform distribution and x_{2i} and x_{3i} are normally distributed with zero mean, unit variance, and correlation coefficient $2^{-1/2}$.

- Model 4: Example 3 in Fan and Zhang (1999)

$$y_i = \sin\left[8\pi(x_{1i} - 0.5)\right] x_{2i} + \Big\{3.5\Big[\exp\left(-(4x_{1i}-1)^2\right) + \exp\left(-(4x_{1i}-3)^2\right)\Big] - 1.4\Big\} x_{3i} + \varepsilon_i$$

All the variables are defined as in Model 3.

For each replication we fit four different models using 500 observations: a STR-Tree model specified with the sequence of LM tests; a regression-tree estimated with CART; a neural network with 10 hidden neurons estimated with Bayesian regularization (MacKay 1992); and MARS (Friedman 1991). For each estimated model, we generate out-of-sample predictions for the remaining 250 observations and we also compute the mean squared errors (MSE). Table 7 reports the median, the MAD, the maximum, and the minimum of the MSEs over 1000 replications.

TABLE 7. Out-of-Sample mean squared error of competing models over 1000 replications.

	Model 1				Model 2			
	Median	MAD	Min.	Max.	Median	MAD	Min.	Max.
STR-Tree/LM	1.97	0.29	1.15	7.75	1.10	0.07	0.86	1.61
CART	2.47	0.18	1.60	3.97	2.46	0.14	1.88	3.07
Neural Network	1.04	0.06	0.77	1.34	1.17	0.07	0.85	1.56
MARS	7.34	0.41	5.55	9.71	1.12	0.07	0.81	1.54
	Model 3				Model 4			
	Median	MAD	Min.	Max.	Median	MAD	Min.	Max.
STR-Tree/LM	1.57	0.10	1.14	2.27	1.90	0.11	1.46	3.04
CART	1.69	0.12	1.22	2.30	2.09	0.15	1.38	3.25
Neural Network	1.58	0.11	1.16	2.26	1.91	0.14	1.38	3.12
MARS	1.60	0.11	1.16	2.33	1.93	0.14	1.42	3.25

Analyzing the results in Table 7, the STR-Tree model performs quite well. In three out of four cases, the STR-Tree model delivers the lowest median of the out-of-sample MSEs. Only for Model 1, the STR-Tree specification is worse than the neural network, but it is significantly better than the CART and the MARS alternatives.

7. Real Examples

In this section we apply the STR-Tree model to several datasets.

- Boston Housing – Housing values in 506 census tracts of Boston. This is the same dataset used in Breiman, Friedman, Olshen, and Stone (1984).
- Cpus data – The Cpus data is discussed in Venables and Ripley (2002). The goal is to explain the performance of 209 different CPUs by some hardware characteristics.
- Car sales in USA in 1993 – The data were taken from MASS library in R and describe the prices and other 25 variables of 93 new cars models.
- Auto imports – This dataset was taken from Ward's 1985 Automotive Yearbook and consists of 195 prices of cars followed by some features such as: fuel consumption, length, width, engine size, among others.
- Abalone data – This is a dataset originated from Biology and the objective is to predict the age of an abalone from a set of physical measurements. There are 4177 cases and 7 continuous predictors. The source is the UCI repository.
- MPG data – The dataset concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 5 continuous attributes. There are 398 observations.

By choosing the datasets above we consider both small and large samples. In some cases the regressors are highly correlated. In all cases we select only the continuous variables. To get an honest picture of the performance reached by all models, we conduct an out-of-sample evaluation by repeating 10 times a 10-fold cross-validation experiment. In each of the 10 replications we randomly split the data in 10 parts, using nine parts to estimate the model and one part to evaluate the out-of-sample performance. We repeat this leaving each one of the 10 parts for out-of-sample evaluation. This means that for each of the 10 replication, we have 10 sets of mean squared errors with $N/10$ observations in each set. N is the number of observations in the dataset. As we repeat the experiment 10 times, in the end we have $10N$ out-of-sample squared errors, reflecting different combinations of estimation (in-sample training) and testing (out-of-sample evaluation) sub-samples. Table 8 reports the median, the MAD, the maximum, and the minimum of the squared errors.

We compared the performance of the following models: CART, MARS, STR-Tree specified with the sequence of LM tests (STR-Tree/LM), STR-Tree specified with 10-fold cross-validation (STR-Tree/CV), and a Neural Network with 10 hidden units estimated with Bayesian regularization. We also consider three possible combination models using a simple averaging scheme. The combinations are: MARS and CART, MARS and STR-Tree/LM, and CART and STR-Tree/LM.

TABLE 8. Out-of-Sample squared errors of different models.

	Boston				Cpus			
Model	Median	MAD	Min.	Max.	Median	MAD	Min.	Max.
CART	20.49	6.99	8.35	51.78	5.56×10^3	4.10×10^3	472.50	4.31×10^4
MARS	11.71	2.91	6.30	39.05	2.34×10^3	1.27×10^3	510.49	1.43×10^4
NN	12.05	4.00	3.85	40.01	2.65×10^3	1.73×10^3	378.37	5.17×10^4
STR-Tree/LM	13.91	3.34	5.64	41.03	2.56×10^3	1.33×10^3	552.19	1.81×10^4
STR-Tree/CV	12.06	2.96	6.49	43.32	3.05×10^3	1.94×10^3	280.00	2.67×10^4
MARS + CART	13.15	4.45	5.07	34.29	2.99×10^3	1.64×10^3	212.72	2.21×10^3
MARS + STR-Tree/LM	**10.38**	3.38	5.08	31.02	$\mathbf{2.08 \times 10^3}$	1.05×10^3	476.43	1.25×10^3
CART + STR-Tree/LM	14.27	3.97	5.24	40.20	3.14×10^3	1.95×10^3	403.72	2.59×10^3
	Car Sales				Import			
Model	Median	MAD	Min.	Max.	Median	MAD	Min.	Max.
CART	33.63	14.85	4.07	229.24	8.14	2.48	2.47	20.24
MARS	26.30	11.73	4.65	156.20	10.42	4.30	2.89	78.20
NN	48.09	27.79	5.26	627.95	14.29	9.43	1.65	112.38
STR-Tree/LM	25.58	12.79	2.97	169.22	8.92	2.38	3.50	26.00
STR-Tree/CV	26.40	15.68	3.08	169.66	11.27	3.05	3.94	33.32
MARS + CART	26.76	12.09	2.58	168.98	**6.35**	2.33	1.91	26.07
MARS + STR-Tree/LM	**22.49**	10.92	4.22	161.24	8.32	2.53	2.50	33.26
CART + STR-Tree/LM	25.38	13.29	3.08	176.48	6.47	2.06	2.58	22.07
	Abalone				MPG			
Model	Median	MAD	Min.	Max.	Median	MAD	Min.	Max.
CART	5.93	0.45	4.54	8.20	13.50	3.17	4.33	23.71
MARS	**4.50**	0.40	3.62	5.84	8.09	1.80	3.77	16.32
NN	4.60	0.47	3.32	7.55	9.09	2.90	3.33	27.03
STR-Tree/LM	5.23	0.55	3.85	7.79	9.54	2.18	4.45	22.96
STR-Tree/CV	6.26	0.63	4.21	8.38	8.06	2.04	2.96	17.60
MARS + CART	4.88	0.41	3.70	6.50	9.00	1.93	3.41	16.69
MARS + STR-Tree/LM	4.79	0.41	3.61	5.96	**7.96**	1.82	3.58	15.45
CART + STR-Tree/LM	5.26	0.44	3.86	6.79	9.86	2.44	3.71	21.63

When compared to CART, the STR-Tree/LM model has a better performance in five out of six cases. The only exception is the Auto Imports dataset, where CART performs slightly better. The STR-Tree/LM model outperforms the NN alternative in three out of six cases (CPUS, Car sales, Import). In the MPG dataset the STR-Tree/LM is worse than the NN model, but the STR-Tree/CV has a lower median of the MSEs. Comparing with MARS, the STR-Tree/LM model has a superior behavior in two out of six cases (Car sales and Import). The STR-Tree/CV is better than the MARS in the MPG dataset. With respect to the Boston dataset, the STR-Tree/CV, NN, and CART models have similar performance. When the CPUS dataset is considered, the out-of-sample behavior of the STR-Tree/LM, NN, and CART specifications are similar. Finally, the simple averaging of the STR-Tree/LM and MARS models leads to the best alternatives in four out of six cases. The two exceptions are the Abalone and the Import datasets. In the former MARS is the best model while in the latter the combination of MARS and CART turns to be the best alternative.

8. Conclusions

In this paper, we proposed a model that combines regression trees and smooth transition regressions. The model is called the Smooth Transition Regression Tree (STR-Tree). The resulting model can be analyzed as a smooth transition regression with multiple regimes. A detailed analysis of the asymptotic properties of the parameter estimates was presented and a model building procedure, based on a sequence of Lagrange Multiplier (LM) tests, was developed. An alternative specification strategy based on a 10-fold cross-validation was discussed and a Monte Carlo experiment was carried out to evaluate the performance of the proposed methodology. The STR-Tree model outperforms CART when the correct selection of the architecture of simulated trees is considered. Furthermore, the LM test seems to be a promising alternative to 10-fold cross-validation. In addition, the proposed estimation algorithm works properly in small samples. When put into proof with real datasets, the STR-Tree model outperformed CART and was highly competitive against other nonlinear alternatives.

References

AHN, H. (1996): “Log-Gamma Regression Modeling Through Regression Trees,” *Communications in Statistics – Theory and Methods*, 25, 295–311.

AMEMIYA, T. (1983): “Non-Linear Regression Models,” in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D.Intriligator, vol. 1, pp. 333–389. Elsevier Science.

BENJAMINI, Y., AND Y. HOCHBERG (1995): “Controlling the False Dicovery Rate - A practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society – Series B*, 57, 289–300.

——— (1997): “Multiple Hypotheses Testing with Weights,” *Scandinavian Journal of Statistics*, 24, 407–418.

BENJAMINI, Y., AND W. LIU (1999): “A Step-Down Multiple Hypothesis Testing Procedures that Controls the False Discovery Rate Under Independence,” *Journal of Statistical Inference and Planning*, 82, 163–170.

BENJAMINI, Y., AND D. YEKUTIELI (2000): “On the Adaptive Control of the Discovery Fate in Multiple Testing with Independent Statistics,” *Journal of Educational and Behavioral Statistics*, 25, 60–83.

——— (2001): “The Control of the False Discovery Rate in Multiple Testing Under Dependency,” *Annals of Statistics*, 29, 1165–1188.

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE (1984): *Classification and Regression Trees*. Belmont Wadsworth Int. Group, New York.

CHAN, K. S., AND H. TONG (1986): “On Estimating Thresholds in Autoregressive Models,” *Journal of Time Series Analysis*, 7, 179–190.

CHANG, R., AND T. PAVLIDIS (1977): “Fuzzy Decision Tree Algorithms,” *IEEE Transactions on Systems, Man, and Cybernetics*, 7, 28–35.

CIAMPI, A. (1991): “Generalized Regression Trees,” *Computational Statistics and Data Analysis*, 12, 57–78.

COOPER, S. J. (1998): “Multiple Regimes in US Output Fluctuations,” *Journal of Business and Economic Statistics*, 16(1), 92–100.

CROWLEY, J., AND M. L. BLANC (1993): “Survival Trees by Goodness of Split,” *Journal of the American Statistical Association*, 88, 457–467.

DAVIES, R. B. (1977): “Hypothesis Testing When the Nuisance Parameter in Present Only Under the Alternative,” *Biometrika*, 64, 247–254.

——— (1987): “Hypothesis Testing When the Nuisance Parameter in Present Only Under the Alternative,” *Biometrika*, 74, 33–44.

DENISON, T., B. K. MALLIK, AND A. F. M. SMITH (1998): “A Bayesian CART Algorithm,” *Biometrika*, 85, 363–377.

EITRHEIM, ., AND T. TERÄSVIRTA (1996): “Testing the Adequacy of Smooth Transition Autoregressive Models,” *Journal of Econometrics*, 74, 59–75.

FAN, J., AND W. ZHANG (1999): “Statistical Estimation in Varying-Coefficient Models,” *Annals of Statistics*, 27, 1491–1518.

FRIEDMAN, J. H. (1991): “Multivariate Adaptive Regression Splines,” *Annals of Statistics*, 19(1), 1–142.

GRANGER, C. W. J., AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

HANSEN, B. E. (1996): “Inference when a Nuisance Parameter is not Identified Under the Null Hypothesis,” *Econometrica*, 64, 413–430.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

HOCHBERG, Y. (1988): “A Sharper Bonferroni Procedure for Multiple Tests of Significance,” *Biometrika*, 75, 800–802.

JAJUGA, K. (1986): “Linear Fuzzy Regression,” *Fuzzy Sets and Systems*, 20, 343–353.

JANG, J. (1994): “Structure Determination in Fuzzy Modeling: A Fuzzy CART Approach,” in *Proceedings of the IEEE Conference on Fuzzy Systems*, pp. 480–485.

JANICKOW, C. (1998): “Fuzzy Decision Trees: Issues and Methods,” *IEEE Transactions on Systems, Man, and Cybernetics*, 28, 1–14.

JENNRICH, R. I. (1969): “Asymptotic Properties of Non-linear Least Squares Estimators,” *The Annals of Mathematical Statistics*, 40, 633–643.

LEWIS, P. A. W., AND J. G. STEVENS (1991): “Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines,” *Journal of the American Statistical Association*, 86, 864–877.

LUUKKONEN, R., P. SAIKKONEN, AND T. TERÄSVIRTA (1988): “Testing Linearity Against Smooth Transition Autoregressive Models,” *Biometrika*, 75, 491–499.

MACKAY, D. J. C. (1992): “A practical Bayesian framework for Backpropagation Networks,” *Neural Computation*, 4, 448–472.

MEDEIROS, M. C., T. TERÄSVIRTA, AND G. RECH (2006): “Building Neural Network Models for Time Series: A Statistical Approach,” *Journal of Forecasting*, 25, 49–75.

MEDEIROS, M. C., AND A. VEIGA (2005): “A Flexible Coefficient Smooth Transition Time Series Model,” *IEEE Transactions on Neural Networks*, 16, 97–113.

MORGAN, J., AND J. SONQUIST (1963): "Problems in The Analysis of Survey Data and a Proposal," *Journal of the American Statistical Association*, 58, 415–434.

MURTHY, S. K. (1998): "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey," *Data Mining and Knowledge Dicovery*, 2, 345–389.

OLARU, C., AND L. WEHENKEL (2003): "A Complete Fuzzy Decision Tree Technique," *Fuzzy Sets and Systems*, 138, 221–254.

SEGAL, M. R. (1992): "Tree-Structured Methods for Longitudinal Data," *Journal of the American Statistical Association*, 87, 407–418.

SUÁREZ, A., AND J. LUTSKO (1999): "Globally Optimal Fuzzy Decision Trees for Classification and Regression," *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 21, 1297–1311.

SUÁREZ, A., AND J. F. LUTSKO (1989): "Tree-Structured Methods for Longitudinal Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, 1297–1311.

TERÄSVIRTA, T. (1994): "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association*, 89, 208–218.

TERÄSVIRTA, T., AND I. MELLIN (1986): "Model Selection Criteria and Model Selection Tests in Regression Models," *Scandinavian Journal of Statistics*, 13, 159–171.

VENABLES, W. N., AND B. D. RIPLEY (2002): *Modern Applied Statistics with S*. Springer, New York.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25.

——— (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York, NY.

WOOLDRIDGE, J. M. (1991): "On the application of robust, regression-based diagnostics to models of conditional means and conditional variances," *Journal of Econometrics*, 47, 5–46.

——— (1994): "Estimation and Inference for Dependent Process," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, pp. 2639–2738. Elsevier Science.

YUAN, Y., AND M. SHAW (1995): "Induction of Fuzzy Decision Trees," *Fuzzy Sets and Systems*, 69, 125–139.

ZADEH, L. (1965): "Fuzzy Sets," *Information and Control*, 8, 338–353.

Appendix A. PROOFS

Appendix A.1. **Proof of Theorem 1.** Lemma 2 of Jennrich (1969) shows that the conditions (1)–(3) in Theorem 1 are enough to guarantee the existence (and measurability) of the NLSE.

Condition (3) in Theorem 1 is satisfied by Assumption 2. It is easy to prove that $H\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ is continuous w.r.t the parameter vector $\boldsymbol{\psi}$. This follows from the fact that, for each value of $\mathbf{x}_t$, $B_k\left(\mathbf{x}_t;\boldsymbol{\theta}_k\right)$ depends continuously on $\boldsymbol{\theta}_k$, $k=1,\ldots,K$. Similarly, $H\left(\mathbf{x}_t,\boldsymbol{\psi}\right)$ is continuous in $\mathbf{x}_t$, and therefore measurable, for each fixed value of $\boldsymbol{\psi}$. Thus (1) and (2) are satisfied.

Q.E.D

Appendix A.2. **Proof of Theorem 2.** Following Jennrich (1969) and Amemiya (1983), $\boldsymbol{\psi} \overset{a.s.}{\rightarrow} \boldsymbol{\psi}^*$ if the following conditions hold: (1) the parameter space $\boldsymbol{\Psi}$ is compact; (2) $Q_T(\boldsymbol{\psi})$ is continuous in $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ for all $\mathbf{x}_t \in \mathbb{X}$ and for all $y_t \in \mathbb{R}$ and $Q_T(\boldsymbol{\psi})$ is a measurable function of $\mathbf{x}_t$ and y_t for all $\boldsymbol{\psi} \in \boldsymbol{\Psi}$; and (3) $Q_T(\boldsymbol{\psi}) \overset{a.s.}{\rightarrow} Q(\boldsymbol{\psi}) = E\left(y_t - H(\mathbf{x}_t;\boldsymbol{\psi})\right)^2$.

Condition (1) is satisfied by Assumption 3. Using the results of Theorem 2, Condition (2) is satisfied. To check if Condition (3) is satisfied we will follow the steps in Amemiya (1983). From (7) and (8) we get

$$\begin{aligned}Q_T(\boldsymbol{\psi}) &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 + \frac{2}{T}\sum_{t=1}^{T}\left[H(\mathbf{x}_t;\boldsymbol{\psi}^*) - H(\mathbf{x}_t;\boldsymbol{\psi})\right]\varepsilon_t + \frac{1}{T}\sum_{t=1}^{T}\left[H(\mathbf{x}_t;\boldsymbol{\psi}^*) - H(\mathbf{x}_t;\boldsymbol{\psi})\right]^2\\ &\equiv A_1 + A_2 + A_3.\end{aligned}$$

It is straightforward to see that $A_1 \overset{a.s.}{\rightarrow} \sigma^2$ by the Strong Law of Large Numbers. Under Assumption 3 and the continuity of $H\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ on $\boldsymbol{\Psi}$, Theorem 4 in Jennrich (1969) implies that $A_2 \overset{a.s.}{\rightarrow} 0$.

Now it is sufficient to show that the following condition is satisfied.

(3’) $\frac{1}{T}\sum_{t=1}^{T} H(\mathbf{x}_t;\boldsymbol{\psi}_1)H(\mathbf{x}_t;\boldsymbol{\psi}_2)$ converges uniformly in $\boldsymbol{\psi}_1,\ \boldsymbol{\psi}_2 \in \boldsymbol{\Psi}$.

Assumption 1, and the fact that $H(\mathbf{x}_t;\boldsymbol{\psi}) \leq \widetilde{\beta}$, where $\widetilde{\beta} = \sum_{k=1}^{K}\left|\beta_{K+k-1}\right| < \infty$, Condition (3’) is satisfied; see Jennrich (1969). Finally we have to show the following condition is satisfied.

(3”) $\lim\limits_{T\rightarrow\infty}\frac{1}{T}\sum_{t=1}^{T}\left[H(\mathbf{x}_t;\boldsymbol{\psi}^*) - H(\mathbf{x}_t;\boldsymbol{\psi})\right] \neq 0$ if $\boldsymbol{\psi} \neq \boldsymbol{\psi}^*$.

The above condition is satisfied by Assumption 4, which guarantees that the STR-Tree model is globally identified.

Q.E.D

Appendix A.3. **Proof of Theorem 3.** To prove the asymptotically normality of the NLSE we need the following conditions in addition to the ones stated in the proof of Theorem 2.

(4) The true parameter vector $\boldsymbol{\psi}^*$ is interior to $\boldsymbol{\Psi}$.

(5) The score vector satisfies

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\psi}^*)),$$

where

$$\mathbf{C}(\boldsymbol{\psi}^*) = \lim_{T\to\infty} \mathrm{E}\left[\frac{1}{T}\frac{\partial Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}}\frac{\partial Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'}\right].$$

(6) The Hessian

$$\frac{1}{T}\frac{\partial^2 Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} \xrightarrow{p} \mathbf{D}(\boldsymbol{\psi}^*),$$

where

$$\mathbf{D}(\boldsymbol{\psi}^*) = \lim_{T\to\infty} \mathrm{E}\left[\frac{1}{T}\frac{\partial^2 Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'\partial \boldsymbol{\psi}}\right].$$

Assumption 3 guarantees that Condition (4) is satisfied. In order to check if Condition (5) is satisfied we have to analyze the behavior of

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} = \frac{2}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'}.$$

As, by Assumption 2, $\varepsilon_t \sim \mathrm{N}(0,\sigma^2)$, we have to show that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}}\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} \equiv \mathbf{H}$$

exists and is non-singular; see Amemiya (1983). First, note that

$$\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} = \left(B_1(\mathbf{x}_t;\boldsymbol{\theta}_1^*),\, \ldots,\, B_K(\mathbf{x}_t;\boldsymbol{\theta}_K^*),\, \beta_{K-1}^*\frac{\partial B_1(\mathbf{x}_t;\boldsymbol{\theta}_1^*)}{\partial \theta_1'},\, \ldots,\, \beta_{2K-2}^*\frac{\partial B_K(\mathbf{x}_t;\boldsymbol{\theta}_K^*)}{\partial \theta_1'}\right)'.$$

By the definition of the STR-Tree model, $B_k(\mathbf{x}_t;\boldsymbol{\theta}_k^*) \leq 1$, $k = 1,\ldots,K$. Furthermore $B_k(\mathbf{x}_t;\boldsymbol{\theta}_k^*)$, $k = 1,\ldots,K$, is the product of at most d (depth of the STR-Tree model) logistic functions of $\mathbf{x}_t$, such that

$$\frac{\partial B_k(\mathbf{x}_t;\boldsymbol{\theta}_k^*)}{\partial \theta_k'} \leq a(\mathbf{x}_t;\boldsymbol{\theta}_k^*) + \sum_{j=1}^{d} c_j(\mathbf{x}_t;\boldsymbol{\theta}_k^*)\left|x_{s_{j-1}t}\right|,\;\; k = 1,\ldots,K, \tag{A.1}$$

where $a(\mathbf{x}_t;\boldsymbol{\theta}_k^*) \leq M < \infty$ and $c_j(\mathbf{x}_t;\boldsymbol{\theta}_k^*) \leq 1$, $j = 1,\ldots,d$. Then, Assumption 2, the unique identification of $\boldsymbol{\psi}^*$ (Assumption 5), and (A.1) guarantee that Condition (5) is satisfied.

To verify Condition (6) we have to show that:

(6') The sum

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'}$$

converges uniformly in $\boldsymbol{\psi}$ in an open neighborhood of $\boldsymbol{\psi}^*$.

(6") The sum

$$\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial^2 H(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right]^2$$

converges uniformly in $\boldsymbol{\psi}$ in an open neighborhood of $\boldsymbol{\psi}^*$.

First, $H(\mathbf{x}_t;\boldsymbol{\psi}^*)$ is twice continuously differentiable and following the same reasoning as before

$$\frac{\partial^2 B_k(\mathbf{x}_t;\boldsymbol{\theta}_k^*)}{\partial\theta_k\partial\theta_k'} \leq u(\mathbf{x}_t;\boldsymbol{\theta}_k^*)+\sum_{i=1}^{d}\sum_{j=1}^{d}v_{ij}(\mathbf{x}_t;\boldsymbol{\theta}_k^*)\left|x_{s_{i-1}t}\right|\left|x_{s_{j-1}t}\right|,\ \ k=1,\ldots,K, \tag{A.2}$$

where $u(\mathbf{x}_t;\boldsymbol{\theta}_k^*) \leq M' < \infty$ and $v_{ij}(\mathbf{x}_t;\boldsymbol{\theta}_k^*) \leq 1$, $j=1,\ldots,d$. Then Condition (6") is satisfied.

Q.E.D

Appendix A.4. **Proof of Theorem 4.** This is a standard result in regression analysis and the proof will be thus omitted.

Q.E.D

(J. C. da Rosa) DEPARTMENT OF STATISTICS, FEDERAL UNIVERSITY OF PARANÁ, CURITIBA, PR, BRAZIL.

E-mail address: joelm@est.ufpr.br

(A. Veiga) DEPARTMENT OF ELECTRICAL ENGINEERING, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: alvf@ele.puc-rio.br

(M. C. Medeiros) DEPARTMENT OF ECONOMICS, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: mcm@econ.puc-rio.br