mathematics

MDPI

Article

CART Leaf Summaries via Bootstrap Bias Correction

George Ekow Quaye [1,*], John Koomson [2], Kwesi Appau Ohene-Obeng [3] and Xiaogang Su [3]

[1] Biostatistics and Computational Biology Core, Center for Population Health and Outcomes Research, Children's Mercy Kansas City, Kansas City, MO 64108, USA
[2] Department of Mathematical Sciences, Central Connecticut State University, New Britain, CT 06050, USA; jkoomson@ccsu.edu
[3] Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA; kaoheneoben@miners.utep.edu (K.A.O.-O.); xsu@utep.edu (X.S.)
* Correspondence: gequaye@cmh.edu

Abstract

CART provides interpretable tree models, but its data-adaptive construction can distort inference for terminal-node summaries. Comparing fitted trees with large independent reference samples, we find that regression terminal-node means are generally well calibrated, whereas standard deviations tend to be underestimated; in binary classification, terminal-node class proportions are often overly extreme, leading to downward-biased Bernoulli variance estimates. We propose a bootstrap bias correction (BBC) procedure that accounts for variation in tree structure by matching bootstrap-tree nodes to those of the original tree using observation overlap. Across simulations, BBC reduces error in regression standard deviations under nonlinear signal structures and improves class-probability and variance estimation in classification. Its benefits diminish, and it may slightly overcorrect, when the true tree structure is recovered nearly perfectly and uncorrected error is already small. Applications to Major League Baseball salary and Pima diabetes data illustrate BBC as a practical approach for mitigating adaptive distortion in CART terminal-node estimation.

Keywords: bootstrap bias correction; CART; post-selection inference; terminal-node estimation; overlap mapping; nonparametric bootstrap

MSC: 62G09; 62G08

1. Introduction

Decision trees are widely used for regression and classification because they accommodate complex relationships between predictors and outcomes while retaining transparent, subgroup-based interpretations [1–3]. CART recursively partitions the predictor space into terminal regions, within which simple response summaries support prediction and scientific description.

Inference for terminal-node summaries is more difficult than their interpretation suggests. The observations used to estimate a node-specific quantity also determine the sequence of splits and, hence, the terminal region itself. Terminal nodes are therefore data-adaptive regions rather than prespecified groups, and analyses that treat the fitted partition as fixed can be overly optimistic. This is a form of post-selection inference, for which conventional procedures need not retain their nominal operating characteristics [4,5]; related work has considered tree-selected subgroups and selective inference for CART [6,7]. In regression trees, adaptive splitting may make training-sample within-node standard deviations smaller than the variability among new observations in the same fitted region, yielding

Check for updates

Academic Editor: Leonid V. Bogachev

Received: 15 September 2026
Revised: 18 October 2026
Accepted: 22 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

overly narrow intervals for terminal-node means. In classification trees, it may produce class-probability estimates that are more extreme than their counterparts in independent observations and, consequently, understate Bernoulli variability.

To characterize these effects, we use terminal-node quantities estimated from a large independent reference sample as a practical gold standard for the corresponding quantities within the fitted terminal regions. For each tree fitted to a training sample, the reference observations are routed through that same tree, and their terminal-node estimates are compared with estimates calculated from the training observations. Because the fitted partition is held fixed and the reference sample is independent of tree construction, this comparison directly assesses the discrepancy associated with using the training data both to select the tree and to estimate its terminal-node quantities. Across the settings studied, terminal-node means in regression trees agree closely with their independent-sample counterparts, whereas terminal-node standard deviations tend to be underestimated. In classification trees, training-sample class proportions tend to be overly extreme, with corresponding downward distortion in estimated Bernoulli variances.

Motivated by these findings, we develop a tree-specific modification of bootstrap bias correction (BBC; [4,8]) for terminal-node estimation. A conventional node-wise BBC requires estimates that are directly comparable across bootstrap replicates and the original sample. For CART, however, a direct node-wise application is not well defined: refitting the tree to each bootstrap sample can alter the split variables, cutpoints, terminal-node regions, and number of terminal nodes, leaving no one-to-one correspondence between bootstrap and original-tree nodes. Our central methodological contribution is an overlap-weighted node-mapping step. Within each bootstrap tree, we compare the terminal-node quantity calculated from the bootstrap observations with the corresponding quantity calculated after the original observations are routed through that same tree. We then aggregate these bootstrap-node discrepancies for each original terminal node according to the overlap of observation memberships. This modification accommodates structural variation across resamples while producing node-specific corrections for the original fitted tree. For regression trees, BBC corrects terminal-node standard deviations directly; for classification trees, it corrects terminal-node class probabilities, from which corrected Bernoulli variances are obtained.

This article makes two primary methodological contributions. First, we characterize how adaptive tree construction affects terminal-node quantities in regression and binary classification trees, motivating response-specific corrections. Second, we develop an overlap-mapped bootstrap framework that transfers bootstrap-node discrepancies to the terminal nodes of the original fitted tree without requiring bootstrap trees to reproduce the original partition. We evaluate these methodological contributions through simulation studies spanning multiple data-generating mechanisms and sample sizes and illustrate their practical consequences using Major League Baseball salary data and Pima diabetes data.

The remainder of this article is organized as follows. Section 2 reviews CART, post-selection inference for tree models, and related bootstrap approaches. Section 3 develops the proposed BBC methodology. Sections 4 and 5 present the simulation study and real-data applications, respectively. Section 6 discusses the findings, limitations, and directions for future work.

2. Background and Related Work

This section provides the conceptual foundation for the proposed procedure. We first introduce CART terminal-node summaries and explain why their data-adaptive construction creates a post-selection inference problem. We then review related selective-inference

and bootstrap ideas, emphasizing why conventional node-wise bias correction does not transfer directly to trees refitted across bootstrap samples.

2.1. Classification and Regression Trees

Let $\mathcal{T}$ denote a fitted CART model constructed from training observations $\{(X_i, Y_i)\}_{i=1}^n$, with terminal nodes $t = 1, \ldots, T$. Terminal node t defines a region $\mathcal{R}_t$ of the predictor space. CART recursively selects binary splits that improve a node-purity criterion and then prunes the resulting tree to control complexity [1]. Each predictor vector is assigned to exactly one terminal region, yielding an interpretable representation of the response distribution within data-defined subgroups.

For a continuous response, the terminal-node sample mean is

$$\bar{Y}_t = \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} Y_i, \tag{1}$$

where n_t is the number of training observations assigned to terminal node t. The corresponding within-node sample standard deviation is

$$S_t = \left\{ \frac{1}{n_t - 1} \sum_{i:X_i \in \mathcal{R}_t} (Y_i - \bar{Y}_t)^2 \right\}^{1/2}. \tag{2}$$

For a binary response $Y \in \{0, 1\}$, the terminal-node class proportion is

$$\hat{p}_t = \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} Y_i, \tag{3}$$

with plug-in Bernoulli variance

$$\hat{V}_t = \hat{p}_t(1 - \hat{p}_t). \tag{4}$$

These quantities have their usual within-group interpretations when the regions $\mathcal{R}_t$ are fixed in advance. In CART, however, the training outcomes influence both the formation of $\mathcal{R}_t$ and the quantities estimated within it. Terminal-node summaries are therefore post-selection quantities rather than summaries from prespecified groups, which motivates the inferential and bootstrap issues considered below.

2.2. Post-Selection Inference for Tree Models

Adaptive tree construction creates a post-selection inference problem: the same data are used both to select the tree and to conduct inference within its selected terminal nodes. Consequently, conventional confidence intervals and hypothesis tests need not retain their nominal properties. Common approaches to post-selection inference include sample splitting, simultaneous inference, and conditional selective inference [5].

This challenge is especially acute for recursive partitioning because selection determines not only the split variables and cutpoints in the fitted tree but also the predictor-space regions within which subsequent inference is performed. A terminal node is therefore a data-selected subgroup rather than a subgroup specified independently of the observed outcome.

In clinical-trial subgroup analysis, Loh et al. [6] emphasized the need to account for post-selection effects in regression-tree-selected subgroups. Their bootstrap-calibration procedure adjusts normal-theory intervals so that expected coverage, averaged across selected subgroups, approximates a specified confidence level.

More recently, Neufeld et al. [7] developed a selective-inference framework for CART regression trees that conditions on the fitted tree having been selected from the data. Under its modeling assumptions, the framework provides tests for differences between a pair of terminal-node means and confidence intervals for a terminal-node mean, demonstrating that naive analyses need not achieve nominal Type I error control or confidence-interval coverage.

The present work addresses a related but distinct question. Rather than deriving a selective distribution conditional on the selected tree, we investigate how adaptive tree construction affects the terminal-node quantities used in conventional node-level analyses. For regression trees, we assess whether the terminal-node standard deviation S_t represents the variability of new observations assigned to the same fitted region. For classification trees, we assess whether the fitted class proportion $\hat{p}_t$ represents the conditional outcome probability within the fitted region and whether $\hat{p}_t(1-\hat{p}_t)$ adequately estimates the corresponding Bernoulli variance. The proposed BBC procedure is intended to reduce adaptive distortion in these terminal-node summaries; it does not provide selective confidence intervals with formal post-selection coverage guarantees.

2.3. Bootstrap Methods for Tree-Based Inference

The nonparametric bootstrap estimates an estimator's sampling variability and bias by repeatedly resampling the observed data [8]. Under standard conditions, the resulting bootstrap distribution approximates the sampling distribution of an estimator without requiring a fully specified parametric model.

For decision trees, resampling changes the fitted tree structure as well as the terminal-node quantities being estimated. A bootstrap tree can differ from the original fitted tree in its split variables, cutpoints, number of terminal nodes, and terminal-node regions. Consequently, terminal nodes from different trees cannot generally be matched by their numerical labels or placed in one-to-one correspondence.

Earlier bootstrap-calibration methods provide a related foundation. Loh [9] and Loh [10] developed procedures for calibrating confidence coefficients, and Loh et al. [6] applied bootstrap calibration to treatment-effect intervals in regression-tree-selected subgroups. These approaches motivate the use of resampling to account for uncertainty introduced by data-dependent tree construction.

Classical bootstrap bias correction provides a complementary perspective by using resampling to estimate systematic estimator bias. Let $\hat{\theta}$ denote an estimator computed from the original sample, and let $\hat{\theta}^*$ denote the corresponding estimator computed from a bootstrap sample drawn with replacement from the original data. The ordinary bootstrap estimate of bias is

$$\widehat{\text{Bias}}(\hat{\theta}) = E_*(\hat{\theta}^*) - \hat{\theta}, \tag{5}$$

where E_* denotes expectation with respect to the bootstrap distribution. Bias-corrected estimators use bootstrap replicates to estimate and adjust for systematic discrepancies between an estimator and its target [4,8]. This formulation presumes that the bootstrap and original-sample estimates are directly comparable estimates of the same target.

For terminal-node estimation, that comparability is not available at the node level because $\hat{\theta}^*$ may be associated with a different terminal-node partition in each bootstrap sample. The procedure must therefore refit the full tree-building process within each resample while also transferring node-specific information back to the terminal nodes of the original fitted tree.

The BBC procedure developed in the next section addresses both requirements. For each bootstrap tree, it compares a terminal-node quantity calculated from the bootstrap observations with the corresponding quantity calculated after the original observations are routed through that same bootstrap tree. The resulting bootstrap-node discrepancies are then mapped to the reporting nodes of the original fitted tree using weights based on overlap of observation memberships. This tree-specific modification permits structural variation across resamples without requiring bootstrap trees to share the original partition.

3. Methodology

This section develops the proposed bootstrap bias correction (BBC) procedure for terminal-node summaries after a CART model has been selected. Let $\mathcal{D} = \{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ denote the training data, let $\mathcal{T}$ denote the fitted tree, and let $\mathcal{L}$ denote its set of terminal nodes. The regions $\mathcal{R}_t$ and terminal-node summaries $\bar{Y}_t$, S_t, $\hat{p}_t$, and $\hat{V}_t$ are defined in Section 2. All BBC corrections are reported for the terminal nodes of the original fitted tree $\mathcal{T}$; bootstrap-tree nodes are used only to estimate those corrections. Because the reference-sample comparisons did not reveal systematic distortion in terminal-node means, BBC targets S_t for regression trees and $\hat{p}_t$ for classification trees.

Our inferential target is conditional on the selected tree $\mathcal{T}$. For each terminal region R_t of $\mathcal{T}$, the target is the corresponding population response standard deviation for regression or event probability for classification among future observations with $\mathbf{X} \in R_t$. We use the term *adaptive distortion* to describe the systematic component of the discrepancy between a terminal-node summary estimated from the data used to construct the fitted tree and this conditional population quantity. The realized discrepancy may also contain ordinary finite-sample variation and therefore should not be attributed entirely to adaptive tree construction.

3.1. Overlap-Mapped Bootstrap Bias Correction

For $b = 1, \ldots, B$, let $\mathcal{D}_b$ be a bootstrap sample drawn with replacement from $\mathcal{D}$, and let $\mathcal{T}_b$ denote the tree fitted to $\mathcal{D}_b$ using a tree-selection rule $\mathcal{A}$. Let $\mathcal{L}_b$ be the terminal-node set of $\mathcal{T}_b$. Each bootstrap terminal node is evaluated twice: first using the bootstrap observations that constructed $\mathcal{T}_b$, retaining bootstrap multiplicities, and then using the original observations routed through the same bootstrap tree, with each original observation used once. The difference between these two quantities is a bootstrap-node discrepancy that reflects adaptive fitting within $\mathcal{T}_b$.

Because $\mathcal{T}_b$ need not share the structure of $\mathcal{T}$, these bootstrap-node discrepancies cannot be assigned directly to the original terminal nodes. Let $\mathcal{I}_t = \{i : \mathbf{X}_i \in \mathcal{R}_t\}$ denote the original-observation indices in terminal node t, and let $\mathcal{I}_{bt'}$ denote the indices of original observations routed to terminal node t' of $\mathcal{T}_b$. For $t \in \mathcal{L}$ and $t' \in \mathcal{L}_b$, define the overlap weight and mapped correction by

$$w_{btt'} = \frac{|\mathcal{I}_t \cap \mathcal{I}_{bt'}|}{n_t}, \qquad \hat{c}_t = \frac{1}{B} \sum_{b=1}^{B} \sum_{t' \in \mathcal{L}_b} w_{btt'} \delta_{bt'}, \tag{6}$$

where $n_t = |\mathcal{I}_t|$ and $\delta_{bt'}$ is an outcome-specific bootstrap-node discrepancy. For each b and t, the weights sum to one over $t' \in \mathcal{L}_b$. Thus, $\hat{c}_t$ aggregates information from the bootstrap-tree nodes according to their overlap with original terminal node t. We call $\hat{c}_t$ a correction, rather than a conventional bias estimate, because its sign is defined as the amount added to the original terminal-node summary. A schematic illustration of the observation-overlap mapping procedure is provided in Appendix F (Figure A1).

The overlap weights provide an empirical approximation to the distribution of bootstrap-node membership conditional on belonging to an original terminal region. Consequently, the mapped discrepancy averages bootstrap-node corrections according to their representation within that region. This approximation is most plausible when discrepancies are reasonably homogeneous across the relevant intersections. When bootstrap partitions differ substantially or discrepancies vary strongly within nodes, overlap weighting need not recover the original-region bias. We therefore examine sensitivity to alternative node mappings, terminal-node size, and bootstrap-tree instability in the simulation study.

3.2. Regression-Tree Standard Deviations

For bootstrap terminal node t', let $S^*_{bt'}$ denote the standard deviation calculated from the bootstrap observations and let $S^0_{bt'}$ denote the standard deviation calculated after the original observations are routed through $\mathcal{T}_b$. We use the one-sided discrepancy

$$\delta^S_{bt'} = \left[S^0_{bt'} - S^*_{bt'}\right]_+, \qquad S^{\mathrm{BBC}}_t = S_t + \hat{c}^S_t, \tag{7}$$

where $[u]_+ = \max(u, 0)$ and $\hat{c}^S_t$ is obtained from Equation (6) using $\delta^S_{bt'}$. This one-sided correction reflects the empirically observed downward distortion in terminal-node standard deviations and prevents BBC from reducing an uncorrected S_t. Terminal-node means are left unchanged.

The positive-part operation in Equation (7) imposes a nonnegative correction motivated by the downward standard-deviation discrepancies observed in the nonlinear regression settings. However, this operation does not distinguish adaptive distortion from ordinary bootstrap sampling variability and can introduce upward bias even when the tree partition is fixed. We therefore evaluate the proposed correction against an untruncated signed alternative and a fixed-tree control in Section 4.3. The computational steps for the proposed regression-tree correction are summarized in Algorithm 1.

Algorithm 1: BBC for regression-tree terminal-node standard deviations

Input: Training data $\mathcal{D}$, original tree $\mathcal{T}$, tree-selection rule $\mathcal{A}$, and number of resamples B

Determine original terminal-node memberships, compute S_t, and set $C_t \leftarrow 0$ for all $t \in \mathcal{L}$;
for $b = 1, \ldots, B$ **do**
  Draw $\mathcal{D}_b$ from $\mathcal{D}$ and fit $\mathcal{T}_b \leftarrow \mathcal{A}(\mathcal{D}_b)$;
  Route $\mathcal{D}$ through $\mathcal{T}_b$ and compute the overlap weights $w_{btt'}$;
  For each $t' \in \mathcal{L}_b$, compute $S^*_{bt'}$, $S^0_{bt'}$, and $\delta^S_{bt'} \leftarrow [S^0_{bt'} - S^*_{bt'}]_+$;
  For each $t \in \mathcal{L}$, update $C_t \leftarrow C_t + \sum_{t' \in \mathcal{L}_b} w_{btt'} \delta^S_{bt'}$;
end
Set $S^{\mathrm{BBC}}_t \leftarrow S_t + C_t/B$ for all $t \in \mathcal{L}$;

Output: BBC-corrected terminal-node standard deviations S^{BBC}_t

3.3. Classification-Tree Probabilities

For bootstrap terminal node t', let $\hat{p}^*_{bt'}$ and $\hat{p}^0_{bt'}$ denote the class-1 proportions calculated from the bootstrap observations and from the routed original observations, respectively. The classification discrepancy is $\delta^p_{bt'} = \hat{p}^0_{bt'} - \hat{p}^*_{bt'}$. Unlike the regression discrepancy, it is not truncated because the required correction can be positive or negative. Let $\Pi_{[0,1]}(u) = \min\{1, \max(0, u)\}$ denote projection onto the probability parameter space. The corrected probability and corresponding Bernoulli variance are

$$\tilde{p}^{\mathrm{BBC}}_t = \Pi_{[0,1]}\left(\hat{p}_t + \hat{c}^p_t\right), \qquad V^{\mathrm{BBC}}_t = \tilde{p}^{\mathrm{BBC}}_t\left(1 - \tilde{p}^{\mathrm{BBC}}_t\right), \tag{8}$$

where $\hat{c}^p_t$ is obtained from Equation (6) using $\delta^p_{bt'}$. BBC therefore corrects the terminal-node probability and derives the Bernoulli variance from that corrected probability, rather than correcting the variance separately. The corresponding procedure for classification trees is summarized in Algorithm 2.

Algorithm 2: BBC for classification-tree terminal-node probabilities

Input: Binary-response training data $\mathcal{D}$, original tree $\widehat{\mathcal{T}}$, tree-selection rule $\mathcal{A}$, and number of resamples B

Determine original terminal-node memberships, compute $\hat{p}_t$, and set $C_t \leftarrow 0$ for all $t \in \mathcal{L}$;

for $b = 1, \ldots, B$ **do**

  Draw $\mathcal{D}_b$ from $\mathcal{D}$ and fit $\widehat{\mathcal{T}}_b \leftarrow \mathcal{A}(\mathcal{D}_b)$;

  Route $\mathcal{D}$ through $\widehat{\mathcal{T}}_b$ and compute the overlap weights $w_{btt'}$;

  For each $t' \in \mathcal{L}_b$, compute $\hat{p}^*_{bt'}$, $\hat{p}^0_{bt'}$, and $\delta^p_{bt'} \leftarrow \hat{p}^0_{bt'} - \hat{p}^*_{bt'}$;

  For each $t \in \mathcal{L}$, update $C_t \leftarrow C_t + \sum_{t' \in \mathcal{L}_b} w_{btt'} \delta^p_{bt'}$;

end

Set $\tilde{p}_t^{\mathrm{BBC}} \leftarrow \Pi_{[0,1]}(\hat{p}_t + C_t/B)$ and $V_t^{\mathrm{BBC}} \leftarrow \tilde{p}_t^{\mathrm{BBC}}(1 - \tilde{p}_t^{\mathrm{BBC}})$ for all $t \in \mathcal{L}$;

Output: BBC-corrected terminal-node probabilities $\tilde{p}_t^{\mathrm{BBC}}$ and Bernoulli variances V_t^{BBC}

3.4. Practical Implementation and Node-Level Intervals

Tree selection is the principal computational cost of BBC. The direct implementation applies the original selection rule, including v-fold cross-validation when used, to every bootstrap sample. A faster alternative grows and prunes each bootstrap tree along its cost-complexity sequence and selects the subtree whose number of terminal nodes is closest to $|\mathcal{L}|$. This matched-size approximation avoids a separate cross-validation loop within every bootstrap replication while preserving the complexity of the original fitted tree. In our simulation experiments, it did not materially degrade BBC performance relative to full bootstrap reselection.

Because BBC estimates an average correction rather than tail quantiles of a bootstrap distribution, it can use a modest number of resamples. We use $B = 50$, which lies within the rough range $25 \leq B \leq 100$ used for related tree-bootstrap bias corrections [11]. This range is a practical guideline rather than a universal guarantee, and larger B may be appropriate when the fitted tree or leaf-specific corrections are especially unstable.

We assessed sensitivity to the number of bootstrap resamples, considering five choices: $B = 25$, 50, 100, 250, and 500. Across regression and classification Models A–C at $n = 500$, aggregate estimation accuracy remained relatively stable, although larger B improved the numerical stability of individual terminal-node corrections. These findings support $B = 50$ as a computationally efficient choice for the settings examined. Detailed results are provided in Appendix A.

The corrected summaries can be inserted into conventional node-level plug-in intervals. For regression trees, a normal-theory interval for the terminal-node mean replaces S_t with S_t^{BBC}; for classification trees, the plug-in Wald-type interval is

$$\tilde{p}_t^{\mathrm{BBC}} \pm z_{1-\alpha/2} \sqrt{\frac{V_t^{\mathrm{BBC}}}{n_t}},$$

where n_t is the terminal-node sample size and the interval endpoints are restricted to $[0, 1]$ where necessary.

As with conventional Wald intervals for binomial proportions, finite-sample coverage may be inadequate, particularly for small terminal nodes or probabilities near the boundaries. To examine the inferential consequences of BBC, we evaluate the empirical coverage and average width of nominal 95% uncorrected and BBC-adjusted intervals for both regression and classification trees. Coverage is assessed relative to independent reference-sample estimates of the population quantities within the selected terminal regions. The simulation results are presented in Section 4.6, with detailed numerical summaries in Appendix H.

These intervals illustrate the practical effect of the correction but should not be interpreted as selective confidence intervals with formally established post-selection coverage.

4. Simulation Study

This section examines both the discrepancy induced by adaptive tree construction and the accuracy of BBC across the full simulation design. We first describe the data-generating mechanisms and evaluation protocol, then use the intermediate sample size $n = 500$ to visualize the training–reference discrepancy, and finally report regression and classification results for $n \in \{100, 500, 1000\}$.

4.1. Simulation Design

We evaluated the effect of adaptive CART construction on terminal-node estimation and assessed whether the proposed bootstrap bias correction (BBC) reduces the resulting discrepancy. In every setting, the five predictors were generated independently from Uniform$(0, 1)$. We considered two smooth nonlinear response surfaces, Models A and B, and a simple tree-structured mechanism, Model C.

$$g_A(\mathbf{X}) = 0.1 \exp(4X_1) + \frac{4}{1 + \exp\{-20(X_2 - 0.5)\}} + 3X_3 + 2X_4 + X_5, \tag{9}$$

$$g_B(\mathbf{X}) = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + X_5, \tag{10}$$

$$g_C(\mathbf{X}) = 2 + 21\{X_1 \le 0.5,\ X_2 \le 0.5\}. \tag{11}$$

For regression, $Y = g_m(\mathbf{X}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and $m \in \{A, B, C\}$. For binary classification, $Y \mid \mathbf{X} \sim \text{Bernoulli}\{\text{expit}(\eta_m(\mathbf{X}))\}$, where $\text{expit}(u) = \{1+\exp(-u)\}^{-1}$, $\eta_A(\mathbf{X}) = g_A(\mathbf{X}) - 6.3$, $\eta_B(\mathbf{X}) = \{g_B(\mathbf{X}) - 12.4\}/4.7$, and $\eta_C(\mathbf{X}) = 31\{X_1 \le 0.5,\ X_2 \le 0.5\} - 1.5$. Models A and B are nonlinear benchmark-type settings related to those commonly used to study flexible regression procedures [12]; Model C has a known three-leaf CART representation.

For each response type, model, and training-sample size $n \in \{100, 500, 1000\}$, we generated $R = 100$ independent data sets. CART models were fitted using cost-complexity pruning and the 1-SE cross-validation rule. For each selected tree $\mathcal{T}$, terminal-node summaries were calculated from the training data and compared with summaries obtained by routing an independent reference sample of size $n_0 = 10{,}000$ through the same fitted tree without refitting. These reference values provide a high-precision, practical gold-standard approximation to terminal-node quantities conditional on the selected partition. BBC used $B = 50$ bootstrap samples in every replication. To assess the precision of the independent-reference estimates, we conducted an additional simulation using reference samples of sizes $n_0 = 10{,}000$ and $n_0 = 100{,}000$, while holding the fitted trees fixed. We examined the distribution of reference observations across terminal nodes and quantified differences in the corresponding reference estimates. The design and results of this assessment are reported in Appendix C.

For regression trees, we compared terminal-node means and standard deviations with their independent-reference counterparts; BBC was applied only to the standard deviations. For classification trees, we compared class probabilities and Bernoulli variances, with the corrected variance calculated from the BBC-corrected probability. MAE and RMSE were computed by pooling terminal-node discrepancies across all replications, so trees with more terminal nodes contribute more estimates to the summaries. Percentage MAE improvements were calculated before rounding. For Model C, we also recorded recovery of the known three-leaf partition, allowing either split order but requiring that one terminal region isolate $\{X_1 \le 0.5,\ X_2 \le 0.5\}$ and that both cutpoints lie within 0.05 of 0.5. Mean terminal-node counts for Models A and B and recovery rates for Model C are reported in the results as structural descriptors, separately from the accuracy tables.

4.2. Reference-Sample Diagnostic at $n = 500$

The full simulation results are reported for $n \in \{100, 500, 1000\}$. To make the mechanism behind the aggregate results concrete, we use the intermediate setting $n = 500$ for the diagnostic display in Figure 1. This setting provides a visually informative intermediate-sample view of data-adaptive partitioning and its associated training-reference discrepancy; the figure is explanatory rather than an additional performance analysis.

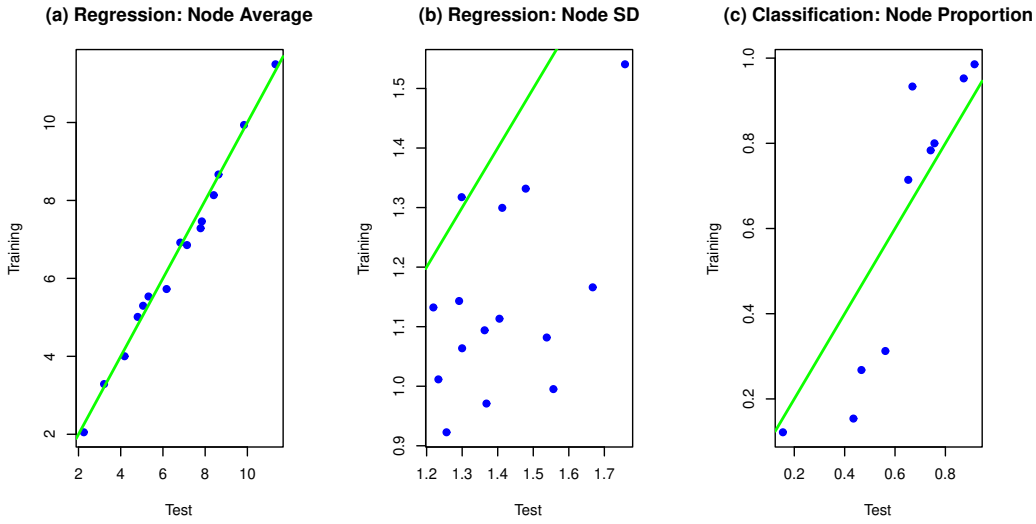


Figure 1. Training-sample and independent-reference terminal-node estimates for a representative Model A data set at the intermediate training-sample size $n = 500$. The panels show regression terminal-node means, regression terminal-node standard deviations, and classification terminal-node probabilities. The horizontal axis gives the independent-reference estimate obtained by routing a sample of size $n_0 = 10{,}000$ through the fitted tree, the vertical axis gives the training-sample estimate, and the blue dots represent individual terminal-node estimates, while the green diagonal line denotes equality. This diagnostic illustration complements the Monte Carlo comparisons across all three training-sample sizes.

Figure 1 illustrates the discrepancy targeted by BBC in an intermediate-sample setting. In this representative replicate, regression terminal-node means lie close to the equality line, whereas most terminal-node standard deviations fall below it. The classification probabilities are also displaced away from 0.5 relative to their independent-reference values. These visual patterns motivate the correction; the tables below assess their frequency and practical consequences across repeated samples and all three training-sample sizes.

4.3. Regression-Tree Results

Table 1 summarizes the regression results. Terminal-node means were strongly concordant with their independent-reference counterparts: correlations ranged from 0.975 to 0.993 across all settings. These correlations support the decision to leave terminal-node means uncorrected, although correlation alone should not be interpreted as a complete assessment of calibration.

The nonlinear regression settings also produced increasingly complex selected trees. The mean (SD) number of terminal nodes for Model A was 4.63 (1.19), 14.93 (3.21), and 19.67 (2.83) at $n = 100$, 500, and 1000, respectively; the corresponding values for Model B were 5.03 (1.20), 16.11 (3.14), and 22.05 (2.74). These structural summaries provide context for the node-level error comparisons reported in Table 1.

Table 1. Accuracy of terminal-node mean and standard-deviation estimates in the regression-tree simulations.

		Terminal-Node Mean			Terminal-Node SD MAE			Terminal-Node SD RMSE	
n	**Model**	**MAE**	**RMSE**	**Corr.**	**Uncorr.**	**BBC**	**Improve. (%)**	**Uncorr.**	**BBC**
100	A	0.358	0.470	0.985	0.312	0.226	27.5	0.381	0.293
	B	0.767	0.964	0.975	0.656	0.433	33.9	0.776	0.548
	C	0.175	0.238	0.973	0.134	0.137	−2.4	0.175	0.181
500	A	0.289	0.382	0.990	0.264	0.160	39.1	0.314	0.209
	B	0.496	0.653	0.990	0.471	0.292	38.0	0.558	0.372
	C	0.070	0.087	0.996	0.050	0.060	−19.8	0.063	0.075
1000	A	0.242	0.324	0.992	0.205	0.132	35.3	0.250	0.176
	B	0.418	0.559	0.993	0.389	0.249	35.9	0.470	0.319
	C	0.050	0.062	0.998	0.036	0.041	−14.0	0.045	0.051

Note. Values pool terminal nodes across $R = 100$ replications. Independent-reference estimates use $n_0 = 10{,}000$ observations, BBC uses $B = 50$ resamples, and only terminal-node SDs are corrected. Positive improvement favors BBC; percentages use unrounded values.

BBC substantially improved terminal-node standard-deviation estimation under the nonlinear Models A and B. For Model A, BBC reduced SD MAE by 27.5% to 39.1% across the three sample sizes; the corresponding reductions for Model B ranged from 33.9% to 38.0%. RMSE also decreased in every nonlinear setting, indicating that the improvement was not driven only by a small subset of terminal nodes.

Model C provided an informative boundary case. True-tree recovery increased from 54% at $n = 100$ to 96% at $n = 500$ and 100% at $n = 1000$, while the uncorrected SD error became small. In contrast to Models A and B, BBC increased both MAE and RMSE for terminal-node standard deviations at every sample size. This pattern is consistent with diminished value of a one-sided SD correction when the fitted partition is recovered reliably and the remaining discrepancy is already small. More generally, it shows that BBC is not uniformly beneficial when little correction is needed.

Sensitivity to Positive-Part Truncation and Fixed-Tree Control

To examine the effect of positive-part truncation, we compared the proposed BBC estimator with an untruncated signed alternative using the same bootstrap samples and overlap weights. The signed correction replaces the positive-part discrepancy in Equation (7) with

$$\delta_{bt'}^{S,\text{signed}} = S_{bt'}^{0} - S_{bt'}^{*}, \qquad S_t^{\text{signed}} = S_t + \frac{1}{B}\sum_{b=1}^{B}\sum_{t' \in \mathcal{L}_b} w_{btt'}\delta_{bt'}^{S,\text{signed}}. \tag{12}$$

We also examined a fixed single-leaf control in which responses were generated independently from $N(0,1)$, with no adaptive partition selection. In this setting, the true population standard deviation is $\sigma = 1$, which serves as the reference value for computing MAE, RMSE, and signed bias. This control allows us to investigate whether positive-part truncation can introduce systematic upward adjustments even in the absence of adaptive tree construction. Both experiments used $R = 100$ Monte Carlo replications and $B = 50$ bootstrap resamples per replication.

Table 2 shows that both BBC variants substantially reduced terminal-node standard-deviation error under the nonlinear Models A and B. Positive-part BBC achieved slightly smaller MAE than signed BBC in five of the six nonlinear settings, although the differences between the two corrected estimators were modest. Both approaches substantially reduced the negative signed bias of the uncorrected estimator.

Table 2. Sensitivity of regression terminal-node standard deviation estimation to positive-part truncation.

		MAE			RMSE			Signed Bias		
n	**Model**	**Uncorr.**	**Signed**	**Positive**	**Uncorr.**	**Signed**	**Positive**	**Uncorr.**	**Signed**	**Positive**
100	A	0.312	0.225	0.226	0.381	0.291	0.293	−0.264	−0.006	0.013
	B	0.656	0.436	0.433	0.776	0.550	0.548	−0.586	−0.065	−0.036
	C	0.134	0.133	0.137	0.175	0.176	0.181	−0.069	0.032	0.050
500	A	0.264	0.161	0.160	0.314	0.210	0.209	−0.243	−0.039	−0.031
	B	0.471	0.294	0.292	0.558	0.374	0.372	−0.431	−0.066	−0.049
	C	0.050	0.056	0.060	0.063	0.069	0.075	−0.011	0.021	0.035
1000	A	0.205	0.133	0.132	0.250	0.177	0.176	−0.183	−0.021	−0.013
	B	0.389	0.252	0.249	0.470	0.322	0.319	−0.359	−0.072	−0.058
	C	0.036	0.038	0.041	0.045	0.048	0.051	−0.004	0.011	0.022
100	Fixed	0.052	0.052	0.056	0.065	0.064	0.070	−0.011	−0.003	0.021
500	Fixed	0.026	0.025	0.027	0.032	0.032	0.036	0.002	0.003	0.015
1000	Fixed	0.018	0.019	0.020	0.023	0.023	0.024	−0.001	−0.001	0.008

Note. Uncorr. denotes the uncorrected terminal-node standard deviation; Signed denotes BBC with untruncated bootstrap discrepancies; and Positive denotes BBC with positive-part truncation. Models A–C use adaptively selected CART trees, with terminal-node estimates compared against independent-reference estimates based on $n_0 = 10{,}000$ observations. The fixed-tree control consists of a single prespecified leaf with responses generated independently from $N(0, 1)$, so that the population standard deviation is one and no adaptive partition selection occurs. Signed bias is the mean estimation error (estimate minus reference); negative and positive values indicate average underestimation and overestimation, respectively. Results are based on $R = 100$ replications and $B = 50$ bootstrap resamples.

Model C exhibited a different pattern. Signed BBC had lower MAE than positive-part BBC at all sample sizes. At $n = 100$, signed BBC marginally improved MAE (0.133 vs. 0.134) but slightly increased RMSE (0.176 vs. 0.175) relative to the uncorrected estimator. At $n = 500$ and $n = 1000$, the uncorrected estimator outperformed both corrections, consistent with improved recovery of the true tree structure. Positive-part BBC also exhibited greater upward bias than signed BBC across all sample sizes.

The fixed-tree control further demonstrates that positive-part truncation can introduce upward bias independently of adaptive partition selection. Its signed bias was 0.021, 0.015, and 0.008 at $n = 100$, 500, and 1000, respectively, compared with values close to zero for the signed alternative. Positive-part BBC also had the largest MAE and RMSE among the three estimators in each fixed-tree setting. Thus, the nonnegative correction can be useful when downward discrepancies are substantial, but its effectiveness is not universal and the resulting adjustment should not be interpreted solely as selection-induced bias.

4.4. Classification-Tree Results

Tables 3 and 4 present the classification results for terminal-node probabilities and Bernoulli variances, respectively. BBC reduced class-probability MAE in every displayed setting. For nonlinear Models A and B, the reductions ranged from 11.0% to 19.0%, and RMSE also declined throughout. BBC produced still larger gains for Bernoulli variance estimation, reducing MAE by 17.4% to 27.2% in Models A and B.

The nonlinear classification trees were smaller but also became more complex as the training sample increased. For Model A, the mean (SD) number of terminal nodes was 2.26 (0.52), 3.98 (2.04), and 4.86 (2.28) at $n = 100$, 500, and 1000, respectively; for Model B, the corresponding values were 2.50 (1.19), 3.66 (2.06), and 4.96 (2.51). These quantities describe

the selected-tree complexity and are therefore reported separately from the probability and variance accuracy measures.

Table 3. Accuracy of terminal-node class-probability estimates in the classification-tree simulations.

		MAE			RMSE	
n	**Model**	**Uncorrected**	**BBC**	**Improvement (%)**	**Uncorrected**	**BBC**
100	A	0.074	0.066	11.0	0.103	0.093
	B	0.149	0.123	17.6	0.174	0.150
	C	0.079	0.068	14.5	0.105	0.091
500	A	0.076	0.063	17.1	0.112	0.097
	B	0.088	0.071	19.0	0.122	0.110
	C	0.027	0.025	7.6	0.034	0.033
1000	A	0.065	0.054	16.8	0.101	0.089
	B	0.076	0.063	16.6	0.107	0.096
	C	0.019	0.018	3.3	0.023	0.023

Note. Values pool terminal nodes across $R = 100$ replications. Errors are calculated against independent-reference probabilities from a sample of size $n_0 = 10{,}000$ routed through each fitted tree. BBC-corrected probabilities are projected onto $[0, 1]$. Improvement is the percentage reduction in MAE; positive values favor BBC. Each setting uses $B = 50$ bootstrap resamples, and percentages were calculated before rounding.

Table 4. Accuracy of terminal-node Bernoulli-variance estimates in the classification-tree simulations.

		MAE			RMSE	
n	**Model**	**Uncorrected**	**BBC**	**Improvement (%)**	**Uncorrected**	**BBC**
100	A	0.040	0.033	17.4	0.053	0.045
	B	0.055	0.041	25.0	0.073	0.057
	C	0.044	0.035	20.0	0.057	0.045
500	A	0.033	0.025	23.5	0.047	0.038
	B	0.026	0.019	27.2	0.038	0.030
	C	0.017	0.016	9.2	0.021	0.020
1000	A	0.027	0.021	22.4	0.040	0.032
	B	0.022	0.017	24.2	0.034	0.028
	C	0.012	0.011	4.0	0.015	0.014

Note. Values pool terminal nodes across $R = 100$ replications. Errors are calculated against independent-reference variances from a sample of size $n_0 = 10{,}000$ routed through each fitted tree. The BBC variance is calculated from the BBC-corrected probability and is not corrected separately. Improvement is the percentage reduction in MAE; positive values favor BBC. Each setting uses $B = 50$ bootstrap resamples, and percentages were calculated before rounding.

The Model C results again show smaller gains as the fitted tree increasingly recovered the known partition. True-tree recovery was 30%, 95%, and 100% at $n = 100$, 500, and 1000, respectively. Probability MAE improvement declined from 14.5% at $n = 100$ to 7.6% at $n = 500$ and 3.3% at $n = 1000$; the corresponding variance-MAE improvements declined from 20.0% to 9.2% and 4.0%. At $n = 1000$, the probability RMSE was unchanged at the displayed precision. Thus, BBC remained beneficial for classification in Model C, but its practical contribution became modest once uncorrected estimates were already accurate.

Taken together, the simulations show that BBC is most useful in settings where adaptive partition selection produces meaningful terminal-node error. It consistently improves the targeted standard-deviation summaries in nonlinear regression settings and improves probability and variance estimation in the classification settings considered. The true-tree

results clarify an important limitation: a positive-only SD adjustment can be unhelpful when the selected tree is already stable and the uncorrected discrepancy is small.

Additional classification simulations examined the sensitivity of BBC to tree-selection rules, class imbalance, and terminal-node size. BBC generally improved estimation in nonlinear Models A and B, although its benefits were less consistent under class imbalance and in Model C. Detailed results are provided in Appendix D.

We also examined BBC under correlated predictors using Gaussian-copula dependence. The correction consistently reduced estimation error for Models A and B across the dependence levels considered, although regression Model C continued to show deterioration. The simulation design and results are presented in Appendix E.

To complement MAE and RMSE, we further examined the direction of terminal-node estimation errors. Appendix G presents signed-error distributions and estimate-versus-reference plots for regression and classification across all models and sample sizes. These results illustrate how BBC reduces underestimation in some settings but may also introduce overcorrection.

We additionally examined the robustness of regression BBC under standardized Student's t_3 and lognormal errors. The simulation design, numerical results, and interpretation are provided in Appendix I.

4.5. Sensitivity to Overlap Mapping and Terminal-Node Stability

We investigated the sensitivity of BBC to the choice of node-mapping strategy. The proposed overlap-weighted correction was compared with a maximum-overlap alternative, which assigns each original terminal node to the bootstrap node having the largest overlap. Both methods used the same bootstrap trees. We considered regression and classification Models A–C at $n = 500$, with $R = 30$ Monte Carlo replications and $B = 50$ bootstrap resamples.

Table 5 shows that both mapping approaches reduced MAE relative to the uncorrected estimator in five of the six settings. Overlap weighting achieved the lowest MAE in regression Models A and B and classification Model C. Maximum-overlap mapping performed best in classification Models A and B, whereas the uncorrected estimator performed best in regression Model C. Thus, neither mapping strategy was uniformly superior.

Table 5. Mean absolute error under alternative node-mapping strategies ($n = 500$, $R = 30$, $B = 50$).

Outcome	Model	Uncorrected	Overlap BBC	Maximum-Overlap BBC
Regression	A	0.2486	**0.1480**	0.1495
	B	0.4910	**0.2950**	0.3027
	C	**0.0503**	0.0558	0.0543
Classification	A	0.0915	0.0779	**0.0745**
	B	0.0817	0.0646	**0.0585**
	C	0.0268	**0.0249**	0.0253

We also examined estimation errors by terminal-node size and bootstrap-tree overlap. Smaller terminal nodes exhibited larger errors in all six settings, whereas the relationship between overlap stability and estimation accuracy was not uniform. Detailed subgroup results are reported in Appendix B, Table A3. These findings highlight the importance of terminal-node size and suggest that overlap alone does not fully characterize the difficulty of terminal-node estimation.

4.6. Empirical Coverage of Confidence Intervals

To investigate the inferential implications of bootstrap bias correction (BBC), we evaluated the empirical coverage probabilities and average widths of nominal 95% confidence intervals for terminal-node estimates. The analysis used the simulation settings described in Section 4.1, covering regression and classification Models A–C with training sample sizes $n = 100$, 500, and 1000. Each setting included $R = 100$ Monte Carlo replications, with $B = 50$ bootstrap resamples used to construct the BBC corrections.

For regression, the conventional interval for the terminal-node mean was constructed using the uncorrected standard deviation, whereas the BBC interval substituted the corrected standard deviation while retaining the original terminal-node mean. For classification, the conventional Wald interval used the uncorrected class proportion and its associated Bernoulli variance, whereas the BBC interval used the corrected class probability and corresponding corrected Bernoulli variance. Classification interval endpoints were restricted to $[0, 1]$.

For each fitted tree, coverage was evaluated against the corresponding terminal-region population quantity approximated by an independent reference sample of size $n_0 = 10{,}000$. Coverage indicators and interval widths were first averaged across terminal nodes within each tree and then across the 100 Monte Carlo replications, thereby assigning equal weight to each fitted tree. Monte Carlo uncertainty was estimated from the tree-level averages because terminal nodes within the same fitted tree are dependent and do not constitute independent simulation replications. With only 100 independent replications per setting, small departures from nominal 95% coverage and small differences between methods should be interpreted cautiously.

Figure 2 shows that BBC generally improved empirical coverage relative to the uncorrected intervals. For regression, coverage increased in all nine simulation settings, with improvements ranging from approximately 1.3 to 7.9 percentage points. For classification, coverage increased in eight of the nine settings and remained unchanged under Model C at $n = 1000$. The largest improvements occurred under classification Model B, where coverage gains ranged from approximately 13.4 to 15.3 percentage points.

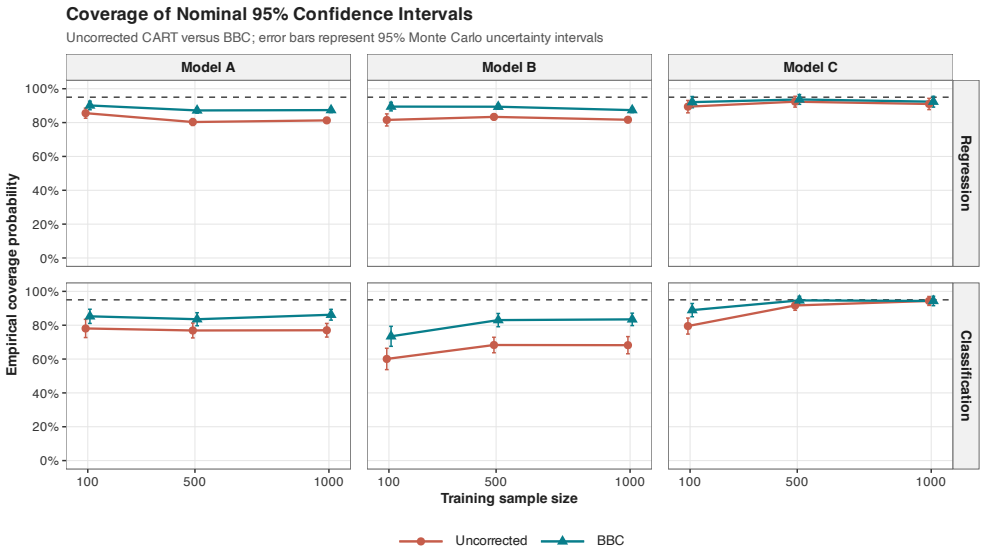


Figure 2. Empirical coverage of nominal 95% confidence intervals for terminal-node estimates under regression and classification Models A–C. The uncorrected CART intervals are compared with their BBC-adjusted counterparts across training sample sizes $n = 100$, 500, and 1000. The horizontal dashed line indicates nominal 95% coverage. Results are based on $R = 100$ Monte Carlo replications per setting.

Despite these improvements, empirical coverage remained below the nominal 95% level in most settings. The BBC adjustment therefore improved the empirical performance of conventional plug-in intervals without consistently achieving nominal coverage.

The coverage gains were accompanied by increases in average interval width. For regression, the relative width increases were more pronounced under Models A and B, ranging from approximately 14.6% to 19.9%. Classification intervals generally exhibited smaller relative increases, ranging from approximately 0.6% to 6.9%. Detailed coverage and interval-width results are reported in Appendix H.

These findings provide additional empirical evidence that correcting adaptive distortion in terminal-node summaries can improve conventional interval performance. However, the intervals do not fully account for uncertainty introduced by adaptive tree selection, and the reference-sample quantities are finite-sample approximations to the corresponding population targets. Consequently, the observed coverage improvements should not be interpreted as formal post-selection coverage guarantees.

5. Real-Data Applications

This section illustrates how overlap-mapped bootstrap bias correction (BBC) changes terminal-node summaries in observed regression and classification data. We analyze a Major League Baseball salary tree and a Pima diabetes tree. Because the corresponding population terminal-node quantities are unknown, these applications demonstrate the direction and practical magnitude of the adjustment rather than its estimation accuracy. In both analyses, CART was selected by cost-complexity pruning with the 1-SE rule, and BBC used $B = 50$ bootstrap resamples.

5.1. Major League Baseball Salary Data

The regression application uses Major League Baseball data containing salary and performance information for 263 players. The response is log salary, with available batting, career, league, division, and fielding characteristics as candidate predictors; variables containing salary information were excluded. The selected four-leaf tree splits on career at-bats (`batcr`), career runs batted in (`rbcr`), and 1986 walks (`wlk86`); it is shown in Figure 3a.

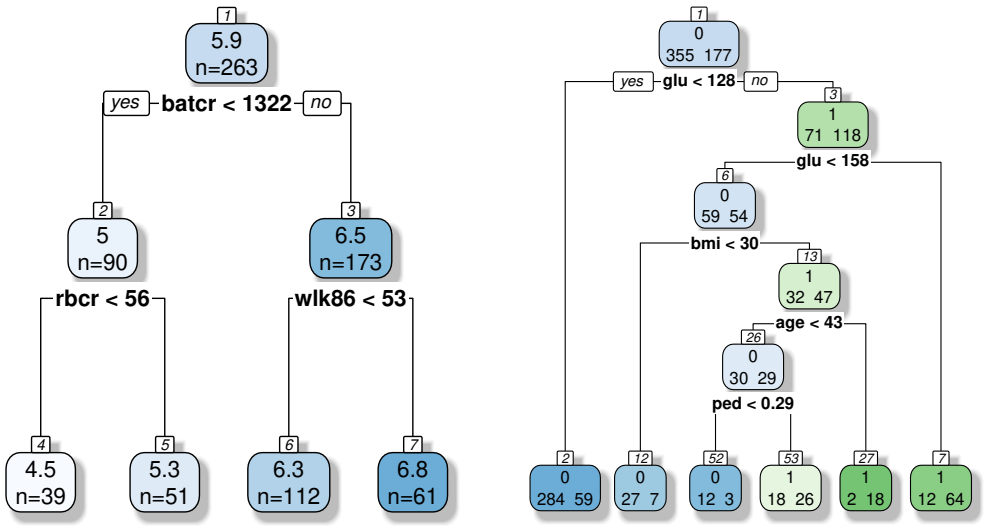


Figure 3. Final 1-SE CART models for the real-data applications: (**a**) the Major League Baseball regression tree and (**b**) the Pima diabetes classification tree.

BBC corrects the terminal-node standard deviation while retaining the fitted node mean. Table 6 reports the uncorrected and BBC-corrected standard deviations and their corresponding plug-in 95% confidence intervals, where "uncorrected" denotes the conventional interval constructed without applying the BBC adjustment.

Table 6. BBC adjustment of terminal-node summaries in the baseball regression tree.

Leaf	n_t	Mean	Uncorrected S_t	BBC S_t	Change (%)	Uncorrected 95% CI	BBC 95% CI
4	39	4.549	0.244	0.272	11.5	(4.472, 4.625)	(4.463, 4.634)
5	51	5.284	0.314	0.377	20.1	(5.198, 5.370)	(5.180, 5.387)
6	112	6.257	0.496	0.568	14.4	(6.165, 6.349)	(6.152, 6.362)
7	61	6.820	0.486	0.564	16.1	(6.698, 6.942)	(6.678, 6.961)

Note. The response is log salary. Change is the percentage increase from the uncorrected to the BBC-corrected standard deviation. Intervals use the corresponding Uncorrected or BBC standard deviation. Because population terminal-node quantities are unknown, the table illustrates the adjustment rather than its accuracy.

BBC increases the standard deviation in every leaf, by 11.5% to 20.1%; the corrected intervals are therefore wider throughout. This direction agrees with the nonlinear regression simulations, but the unknown population standard deviations preclude a direct accuracy assessment.

5.2. Pima Diabetes Data

The classification application uses the Pima diabetes data in the `MASS` package [13]. The supplied training and test partitions were combined because the objective is terminal-node estimation rather than out-of-sample prediction; diabetes is coded as $Y = 1$. The selected six-leaf tree splits on plasma glucose concentration, body mass index, age, and diabetes pedigree function; it is shown in Figure 3b.

For the Pima tree, BBC is applied to the terminal-node diabetes probability, and the corrected Bernoulli variance is then computed from that probability. Table 7 compares the uncorrected and corrected estimates.

Table 7. BBC adjustment of terminal-node summaries in the Pima diabetes classification tree.

Leaf	n_t	$\hat{p}_t$	$\hat{p}_t^{\mathrm{BBC}}$	Δp	V_t	V_t^{BBC}	Uncorrected 95% CI	BBC 95% CI
2	343	0.172	0.186	+0.014	0.142	0.151	(0.132, 0.212)	(0.145, 0.227)
12	34	0.206	0.234	+0.028	0.163	0.179	(0.070, 0.342)	(0.092, 0.376)
52	15	0.200	0.206	+0.006	0.160	0.164	(0.000, 0.402)	(0.001, 0.411)
53	44	0.591	0.573	−0.018	0.242	0.245	(0.446, 0.736)	(0.427, 0.719)
27	20	0.900	0.868	−0.032	0.090	0.114	(0.769, 1.000)	(0.720, 1.000)
7	76	0.842	0.819	−0.023	0.133	0.148	(0.760, 0.924)	(0.733, 0.906)

Note. $V_t = \hat{p}_t(1 - \hat{p}_t)$, and V_t^{BBC} is calculated from the BBC-corrected probability. $\Delta p = \hat{p}_t^{\mathrm{BBC}} - \hat{p}_t$. Intervals are plug-in Wald-type 95% intervals, restricted to $[0, 1]$ where necessary.

All six corrected probabilities move toward 0.5. The largest change occurs in leaf 27, from 0.900 to 0.868, and its Bernoulli variance increases from 0.090 to 0.114. The BBC variance exceeds its uncorrected counterpart in every leaf, echoing the qualitative pattern in the classification simulations. As in the baseball analysis, unknown population quantities prevent a direct evaluation of estimation accuracy.

6. Discussion

This study examined how adaptive CART construction affects terminal-node summaries and developed a tree-specific bootstrap bias correction (BBC) procedure that accom-

modates structural variation among bootstrap trees. Across the nonlinear settings studied, BBC reduced error in regression terminal-node standard deviations and in classification terminal-node probabilities and Bernoulli variances. Terminal-node means in regression trees remained closely aligned with their independent-reference counterparts and were therefore left uncorrected. Model C qualified these findings: as the known tree was recovered more reliably, uncorrected error became small, the gains from BBC diminished for classification, and the nonnegative regression adjustment could increase error. BBC is consequently most useful when adaptive partition selection remains consequential for the terminal-node quantity of interest.

The additional sensitivity analysis clarifies the role of the one-sided regression correction. Both signed and positive-part BBC substantially improved standard-deviation estimation in the nonlinear Models A and B, with only modest differences between them. However, the positive-part correction produced greater upward bias in Model C and in a fixed-tree control without adaptive partition selection. These findings demonstrate that the positive-part adjustment is an empirically motivated, directional correction rather than a mechanism that exclusively estimates selection-induced bias. Its use is most defensible when there is evidence of meaningful downward distortion; when uncorrected estimates are already accurate, the signed correction or no correction may be preferable.

The essential modification to conventional bootstrap bias correction is the overlap mapping from bootstrap terminal nodes to the terminal nodes of the original fitted tree. Because bootstrap refits can change split variables, cutpoints, and terminal regions, direct node-to-node matching is generally unavailable. Mapping by shared observation membership permits node-specific corrections despite this structural instability. In the baseball and Pima applications, BBC produced adjustments in the directions suggested by the simulations; because their true terminal-node quantities are unknown, these examples illustrate the magnitude and practical implications of the correction rather than establish its accuracy.

The proposed BBC procedure is most naturally interpreted as a bootstrap-based correction for terminal-node summaries conditional on the selected tree rather than as a method with formal selective-inference guarantees. Its effectiveness depends on the bootstrap trees providing informative counterparts to the terminal regions of the original tree, so that the overlap mapping transfers relevant bootstrap discrepancies. Consequently, the correction may be less reliable when the fitted tree is highly unstable, terminal nodes are very small, or bootstrap partitions differ substantially from the original partition. The present results should therefore be interpreted as empirical evidence for the proposed correction rather than as establishing general consistency or post-selection coverage guarantees.

BBC is a point-estimation procedure and does not by itself provide formally selective confidence intervals. A separate empirical examination of Loh's bootstrap calibration (BC) in CART settings, reported by Quaye [14], found that its adjusted confidence coefficient did not consistently attain the benchmark indicated by a large independent reference sample. This evidence is diagnostic rather than a general theoretical verdict on BC, but it reinforces the value of evaluating interval calibration under adaptive partitioning. Establishing coverage properties for intervals constructed from BBC-corrected summaries remains an important theoretical problem.

Tree-Values [7] provides a valuable but complementary selective-inference framework: it conditions on CART selection to form tests and confidence intervals for regression-tree leaf means. Its inferential target, assumptions, and formal guarantee therefore differ from those of BBC, which targets terminal-node variability and classification summaries through ordinary bootstrap refits and overlap-based mapping. BBC is deliberately straightforward to implement, computationally practical under the resampling choices described above, and applicable to both regression and classification trees, but it is not a substitute for formal

selective inference when selective coverage or hypothesis testing is the scientific objective. A numerical head-to-head comparison would require a common estimand, compatible tree-fitting assumptions, and an aligned inferential criterion; it would therefore constitute a separate methodological study rather than an additional simulation within the present point-estimation analysis.

The additional confidence-interval experiments showed that BBC generally improved the empirical coverage of conventional nominal 95% plug-in intervals in both regression and classification settings, although the improvements were accompanied by wider intervals. Coverage nevertheless remained below the nominal level in several settings, underscoring the limitations of bias correction for inference following adaptive tree selection. These findings are empirical and do not establish formal post-selection coverage guarantees.

The additional heavy-tailed regression experiments further examined the robustness of BBC beyond the Gaussian error assumption. Under standardized Student's t_3 errors and centered, variance-standardized lognormal errors, BBC reduced both MAE and RMSE for nonlinear Models A and B across all sample sizes. However, its performance under the simpler tree-structured Model C was mixed, with deterioration at larger sample sizes. These findings suggest that the benefits of BBC extend to the non-Gaussian settings examined, but remain dependent on the underlying regression structure. Detailed results are presented in Appendix I.

In practical applications, exploratory data analysis may reveal substantial skewness or irregularity in a continuous response. In such cases, an appropriate response transformation may be considered before fitting a regression tree, depending on the scientific objective and interpretability of the transformed scale. The additional non-Gaussian simulations nevertheless provide useful evidence on the behavior of BBC when the response is analyzed without such a transformation.

The present results are empirical and limited to the examined CART specifications, data-generating mechanisms, and tuning choices. Future work should establish theoretical properties of the overlap-mapped correction, develop diagnostics for determining when correction is warranted, and study interval procedures with demonstrable coverage. A natural methodological extension is to adapt the overlap-mapped correction to other tree-structured models, including interaction trees for treatment-effect heterogeneity [15,16] and survival trees [17]. Such work would test whether node-overlap mapping can accommodate the distinct splitting criteria and terminal-node targets used beyond standard CART.

Author Contributions: Conceptualization, G.E.Q. and X.S.; methodology, G.E.Q. and X.S.; software, G.E.Q.; validation, G.E.Q. and X.S.; formal analysis, G.E.Q.; investigation, G.E.Q. and X.S.; writing—original draft preparation, G.E.Q.; writing—review and editing, G.E.Q., J.K., K.A.O.-O. and X.S.; supervision, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to this study involving secondary analysis of an existing, publicly available dataset containing no direct personal identifiers. No new human participants were recruited, and no identifiable personal information was accessed or collected.

Informed Consent Statement: Ethical review and approval were waived for this study due to the authors used an existing, publicly available dataset and did not recruit participants or obtain new data directly from individuals.

Data Availability Statement: The original contributions presented in this study are included in the article. For further inquiries, please contact the corresponding author.

Acknowledgments: The authors gratefully acknowledge helpful discussions and feedback that contributed to the development and refinement of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Sensitivity to the Number of Bootstrap Resamples

Appendix A.1. Regression Sensitivity

We examined the sensitivity of the regression BBC estimator to the number of bootstrap resamples, considering $B \in \{25, 50, 100, 250, 500\}$ under Models A–C at $n = 500$. The analysis used $R = 30$ Monte Carlo replications, with nested bootstrap sequences to facilitate direct comparisons across values of B.

Table A1. Sensitivity of regression BBC to the number of bootstrap resamples ($n = 500$, $R = 30$).

Model	B	MAE	RMSE	Signed Bias
A	25	0.1616	0.2115	−0.0240
	50	0.1609	0.2103	−0.0257
	100	0.1612	0.2106	−0.0263
	250	0.1607	0.2098	−0.0263
	500	0.1606	0.2098	−0.0266
B	25	0.2872	0.3645	−0.0289
	50	0.2842	0.3609	−0.0276
	100	0.2870	0.3633	−0.0285
	250	0.2860	0.3625	−0.0296
	500	0.2861	0.3624	−0.0299
C	25	0.0656	0.0789	0.0383
	50	0.0664	0.0801	0.0400
	100	0.0662	0.0797	0.0401
	250	0.0662	0.0800	0.0405
	500	0.0659	0.0796	0.0406

Table A1 shows that MAE, RMSE, and signed bias remained stable across bootstrap counts, with no systematic improvement in aggregate accuracy beyond $B = 50$. However, larger B improved node-specific numerical stability, particularly under Model B. Thus, $B = 50$ provides a reasonable computational compromise for the settings examined.

Appendix A.2. Classification Sensitivity

We extended the bootstrap-count sensitivity analysis to classification Models A–C at $n = 500$, using $R = 30$ Monte Carlo replications and $B \in \{25, 50, 100, 250, 500\}$. The same nested bootstrap sequences were used to compare terminal-node probability and Bernoulli-variance estimation across bootstrap counts.

Table A2. Sensitivity of classification BBC to the number of bootstrap resamples ($n = 500$, $R = 30$).

		Probability		Bernoulli Variance	
Model	B	MAE	RMSE	MAE	RMSE
A	25	0.0441	0.0706	0.0156	0.0211
	50	0.0439	0.0710	0.0155	0.0212
	100	0.0434	0.0703	0.0153	0.0209
	250	0.0433	0.0699	0.0151	0.0205
	500	0.0432	0.0699	0.0151	0.0205

Table A2. *Cont.*

Model	B	Probability		Bernoulli Variance	
		MAE	RMSE	MAE	RMSE
B	25	0.0700	0.1066	0.0224	0.0375
	50	0.0705	0.1066	0.0225	0.0377
	100	0.0702	0.1063	0.0223	0.0374
	250	0.0703	0.1063	0.0224	0.0373
	500	0.0706	0.1066	0.0225	0.0374
C	25	0.0265	0.0343	0.0163	0.0210
	50	0.0259	0.0338	0.0160	0.0208
	100	0.0258	0.0334	0.0159	0.0205
	250	0.0258	0.0333	0.0159	0.0205
	500	0.0258	0.0332	0.0159	0.0205

Table A2 shows that probability and Bernoulli-variance estimation errors remained stable across bootstrap counts. At $B = 50$, mean absolute differences relative to $B = 500$ ranged from 0.0029 to 0.0037 for probabilities and from 0.0015 to 0.0018 for variances. These results support $B = 50$ as a reasonable computational compromise, although larger values improve node-specific numerical stability.

Appendix B. Sensitivity to Terminal-Node Size and Bootstrap-Tree Overlap

We further examined the accuracy of overlap-weighted BBC across terminal-node size and bootstrap-tree overlap. Within each outcome and model, terminal nodes were classified using the median original leaf size and, separately, the median mean dominant-overlap score across bootstrap trees. The analysis used $n = 500$, $R = 30$ Monte Carlo replications, and $B = 50$ bootstrap resamples.

Table A3. Mean absolute error of overlap-weighted BBC stratified by terminal-node size and bootstrap-tree overlap.

Outcome	Model	Leaf Size		Bootstrap Overlap	
		Smaller	Larger	Lower	Higher
Regression	A	0.1860	0.1096	0.1343	0.1617
	B	0.3505	0.2390	0.2739	0.3160
	C	0.0634	0.0475	0.0555	0.0561
Classification	A	0.1236	0.0322	0.0784	0.0774
	B	0.0960	0.0314	0.0482	0.0814
	C	0.0269	0.0229	0.0261	0.0238

Table A3 shows that smaller terminal nodes had higher estimation errors in all six settings. In contrast, the relationship between bootstrap-tree overlap and estimation accuracy was not uniform. Lower-overlap nodes had smaller MAE in some settings and larger MAE in others. These findings suggest that terminal-node size is an important determinant of estimation accuracy, whereas overlap alone does not fully explain the behavior of BBC corrections.

Appendix C. Precision of Independent-Reference Estimates

We conducted an additional simulation to assess the numerical precision of the independent-reference estimates used to evaluate terminal-node accuracy. For regression and classification Models A–C, we considered training-sample sizes $n = 500$ and $n = 1000$, with $R = 30$ Monte Carlo replications per setting. Each CART tree was fitted using the same 1-SE selection rule as in the main simulations.

For each fitted tree, we generated an independent reference sample of size 100,000 and used its first 10,000 observations as the smaller reference sample. Both samples were routed through the same fitted tree without refitting. Thus, the two reference estimates were based on nested samples and differed only in reference-sample size.

We recorded the number of reference observations in each terminal node and compared the resulting standard deviations for regression and event probabilities for classification.

Table A4 shows that reference-node counts varied substantially across the simulation settings. The smallest reference counts occurred in the nonlinear regression models, particularly at $n = 1000$, where the fitted trees contained more terminal nodes.

Table A4. Reference-node counts and precision of independent-reference estimates.

			Reference-Node Count			
Outcome	**Model**	***n***	**Min.**	**5th pct.**	**Median**	**Mean |Δ|**
Regression	A	500	111	187	567	0.0349
	A	1000	49	126	457	0.0386
	B	500	93	170	487	0.0609
	B	1000	52	123	343	0.0689
	C	500	2181	2346	2532	0.0118
	C	1000	2297	2367	2532	0.0098
Classification	A	500	93	273	1364	0.0106
	A	1000	99	215	1063	0.0111
	B	500	159	276	2222	0.0114
	B	1000	82	255	1716	0.0102
	C	500	2127	2271	2572	0.0052
	C	1000	2340	2402	2536	0.0050

Note. Each setting includes $R = 30$ independently fitted trees. Reference-node counts are based on $n_0 = 10{,}000$ observations. Δ denotes the difference between terminal-node reference estimates obtained using 10,000 and 100,000 observations. Mean $|\Delta|$ is averaged across terminal nodes. For regression, the estimated quantity is the response standard deviation; for classification, it is the event probability. The reference samples are nested, not mutually independent.

For regression Model B at $n = 1000$, the minimum reference-node count was 52, the fifth percentile was approximately 123, and the median was 343.

The mean absolute differences between reference estimates based on 10,000 and 100,000 observations ranged from 0.0098 to 0.0689 for regression standard deviations and from 0.0050 to 0.0114 for classification probabilities.

These results indicate that reference-sample uncertainty is relatively modest in many settings but is not negligible in some smaller nonlinear regression nodes. The reference estimates should therefore be interpreted as Monte Carlo approximations rather than exact population quantities.

For binary classification, the reference Bernoulli variance was calculated as $\hat{p}_{\text{ref}}(1 - \hat{p}_{\text{ref}})$.

Conditional on a fixed terminal region containing m independent reference observations, this estimator satisfies

$$\mathbb{E}[\hat{p}_{\text{ref}}(1 - \hat{p}_{\text{ref}})] = \left(1 - \frac{1}{m}\right) p(1 - p). \tag{A1}$$

Consequently, the plug-in reference variance has a small finite-sample downward bias. The corresponding unbiased estimator is

$$\widehat{V}_{\text{ref,unbiased}} = \frac{m}{m-1}\hat{p}_{\text{ref}}(1-\hat{p}_{\text{ref}}), \qquad m > 1. \tag{A2}$$

In the present experiment, the mean adjustment from this correction ranged from approximately 0.00005 to 0.00028 across the classification settings. Thus, the finite-sample bias of the reference Bernoulli variance was small on average, although reference-sample variability remains relevant for individual terminal nodes. Overall, the additional analysis supports the use of $n_0 = 10{,}000$ as a practical reference sample while identifying limitations in smaller terminal regions.

The experiment evaluates reference-estimation precision but does not directly establish that every reported BBC improvement or method ranking is unchanged when the reference sample is increased.

Appendix D. Sensitivity to Tree Selection, Class Imbalance, and Terminal-Node Size

We conducted an additional simulation to examine the sensitivity of BBC to tree-selection rules, class imbalance, and terminal-node sample sizes. We considered classification Models A–C with training-sample sizes $n = 500$ and $n = 1000$. Each setting included $R = 30$ Monte Carlo replications, $B = 50$ bootstrap resamples, and an independent reference sample of $n_0 = 10{,}000$ observations.

Two tree-selection rules were evaluated: the 1-SE cross-validation rule used in the main simulations and the minimum-cross-validation-error rule, denoted 0-SE. Both rules were applied to the same generated training data within each replication.

In addition to the original classification models, we considered an imbalanced scenario obtained by subtracting 2.5 from the original log-odds of the event probability. Specifically, if $\eta_m(X)$ denotes the original linear predictor for Model m, the modified probability was

$$p_{m,\text{imb}}(X) = \text{expit}\{\eta_m(X) - 2.5\}, \qquad m \in \{A, B, C\}. \tag{A3}$$

This modification preserved the original predictor relationships while reducing the event prevalence. Across the imbalanced settings, the observed event proportions were approximately 0.21, 0.11, and 0.08 for Models A, B, and C, respectively.

Estimation accuracy was assessed using terminal-node mean absolute error (MAE) for event probabilities and Bernoulli variances. The percentage improvement was calculated as

$$\text{Improvement}(\%) = 100\left(1 - \frac{\text{MAE}_{\text{BBC}}}{\text{MAE}_{\text{Uncorrected}}}\right). \tag{A4}$$

Positive values indicate improvement from BBC, whereas negative values indicate increased estimation error. Results were pooled across terminal nodes within each simulation setting.

Table A5 shows that BBC generally reduced probability and variance estimation error in the nonlinear Models A and B under standard classification. For these models, probability-MAE improvements ranged from approximately 13.8% to 21.0%, while variance-MAE improvements ranged from 22.8% to 28.1%. The improvements were observed under both tree-selection rules.

Under class imbalance, BBC continued to improve estimation for Model A, with probability-MAE reductions of approximately 15.8% to 16.6%. However, the results were less consistent for Models B and C. In several settings, the selected trees contained only one terminal node, leaving limited scope for correcting adaptive partitioning effects.

These findings indicate that BBC can remain useful under alternative tree-selection rules and moderate-to-substantial imbalance, but its benefit is not uniform. In particular, a correction may offer little improvement when the selected tree is simple and uncorrected estimation error is already small.

To investigate terminal-node size, we grouped the fitted terminal nodes according to their training-sample counts: small ($n_t < 20$), medium ($20 \leq n_t < 50$), and large ($n_t \geq 50$). We pooled node-level estimation errors across Models A–C and both training-sample sizes within each scenario and pruning rule. These pooled summaries are descriptive because the distribution of models and tree structures differs across size groups.

Table A5. Sensitivity of BBC estimation accuracy to tree-selection rule and class imbalance.

Scenario	Model	n	Rule	Nodes	Event Rate	p imp. (%)	V imp. (%)
Standard	A	500	1-SE	148	0.50	18.6	26.5
		500	0-SE	207	0.50	18.7	26.9
		1000	1-SE	148	0.50	18.8	23.8
		1000	0-SE	241	0.50	16.1	23.7
	B	500	1-SE	107	0.50	21.0	27.9
		500	0-SE	200	0.50	16.0	25.4
		1000	1-SE	158	0.51	20.8	28.1
		1000	0-SE	270	0.51	13.8	22.8
	C	500	1-SE	90	0.35	−1.7	0.1
		500	0-SE	90	0.35	−2.5	−0.3
		1000	1-SE	90	0.34	2.6	3.4
		1000	0-SE	90	0.34	2.8	3.8
Imbalanced	A	500	1-SE	110	0.22	15.9	20.7
		500	0-SE	133	0.22	15.8	23.2
		1000	1-SE	121	0.21	16.6	25.9
		1000	0-SE	181	0.21	16.3	25.7
	B	500	1-SE	30	0.10	−0.7	−0.6
		500	0-SE	34	0.11	6.4	7.9
		1000	1-SE	30	0.11	−1.6	−1.5
		1000	0-SE	35	0.11	0.6	−1.9
	C	500	1-SE	30	0.08	2.1	2.2
		500	0-SE	30	0.08	0.7	1.0
		1000	1-SE	30	0.08	−1.9	−1.8
		1000	0-SE	33	0.08	3.2	5.8

Note. Each setting includes $R = 30$ fitted trees. Nodes denotes the total number of terminal-node records pooled across replications. Event rate is the mean observed event proportion. The 1-SE rule selects a simpler tree within one standard error of the minimum cross-validation error; 0-SE selects the minimum-error tree. p imp. and V imp. denote percentage reductions in MAE for event probabilities and Bernoulli variances, respectively. Negative values indicate deterioration.

Table A6 shows that terminal-node estimation errors were largest in the smallest leaves. Although BBC reduced pooled probability MAE in each displayed size category, the relative improvements were generally smaller for leaves containing fewer than 20 training observations.

The small-node results should be interpreted cautiously, particularly in the imbalanced 1-SE setting, where only four terminal-node records were available. Moreover, pooling across models and sample sizes does not isolate terminal-node size from differences in signal structure or tree selection. These findings therefore describe observed performance rather than establish a universal leaf-size effect.

Overall, the additional simulations support the usefulness of BBC across the two tree-selection rules examined, particularly for nonlinear classification models with appreciable adaptive partitioning. However, class imbalance, small terminal nodes, and simple selected trees can limit the magnitude and consistency of the correction's benefits.

The analysis does not address all possible departures from the original simulation design, including higher-dimensional covariates or alternative response distributions. Correlated predictors are examined separately in Appendix E. Consequently, these findings should be interpreted as targeted sensitivity evidence rather than a general robustness guarantee.

Table A6. BBC performance by terminal-node size across the additional classification simulations.

Scenario	Rule	Leaf Size	Nodes	Uncorrected p MAE	BBC p MAE	p imp. (%)
Standard	1-SE	<20	45	0.2104	0.1877	10.8
		20–49	109	0.1188	0.0962	19.0
		50+	587	0.0417	0.0333	20.1
	0-SE	<20	192	0.2178	0.1994	8.5
		20–49	258	0.1325	0.1081	18.4
		50+	648	0.0493	0.0392	20.6
Imbalanced	1-SE	<20	4	0.2867	0.2648	7.6
		20–49	39	0.1038	0.0890	14.3
		50+	308	0.0266	0.0224	16.0
	0-SE	<20	42	0.2304	0.2100	8.9
		20–49	82	0.1207	0.1014	16.0
		50+	322	0.0289	0.0233	19.4

Note. Leaf size refers to the number of training observations assigned to a terminal node. MAE values are pooled across Models A–C and $n = 500, 1000$. p imp. denotes the percentage reduction in probability MAE. The imbalanced 1-SE small-node group contains only four terminal-node records; its estimate is therefore particularly unstable.

Appendix E. Sensitivity to Correlated Predictors

We conducted additional simulations to examine the sensitivity of BBC to dependence among predictors. We considered regression and classification Models A–C, with training sample sizes $n = 500$ and $n = 1000$. Each setting included $R = 30$ Monte Carlo replications, $B = 50$ bootstrap resamples, and an independent reference sample of $n_0 = 10{,}000$ observations. The fitted trees used the same 1-SE selection rule as in the main simulations.

Predictor dependence was introduced using a Gaussian copula. Specifically, we generated

$$\mathbf{Z} \sim N_p(\mathbf{0}, \Sigma_\rho), \qquad X_j = \Phi(Z_j), \quad j = 1, \ldots, p, \tag{A5}$$

where Φ denotes the standard normal distribution function and Σ_ρ has unit diagonal elements and common off-diagonal elements ρ. We considered $\rho = 0$, 0.5, and 0.8. This construction preserves uniform marginal distributions while introducing dependence among predictors. The parameter ρ represents latent Gaussian correlation rather than the Pearson correlation of the transformed predictors.

For each fitted tree, we compared uncorrected and BBC terminal-node estimates against estimates obtained from the independent reference sample. Accuracy was summarized using terminal-node mean absolute error (MAE). The percentage improvement was defined as

$$\text{Improvement}(\%) = 100\left(1 - \frac{\text{MAE}_{\text{BBC}}}{\text{MAE}_{\text{Uncorrected}}}\right). \tag{A6}$$

Positive values indicate reduced MAE following BBC, while negative values indicate deterioration.

Table A7 shows that BBC consistently reduced standard-deviation MAE for regression Models A and B at all three dependence levels. Improvements ranged from approximately 36.1% to 44.9% across the examined settings.

In contrast, BBC increased MAE for regression Model C at every dependence level. The deterioration ranged from approximately 0.8% to 20.1%. This result is consistent with the limited benefit of bias correction when the uncorrected terminal-node estimates are already relatively accurate.

Table A7. Sensitivity of regression terminal-node standard-deviation estimation to predictor dependence.

		$n = 500$			$n = 1000$		
Model	ρ	**Uncorrected MAE**	**BBC MAE**	**Imp. (%)**	**Uncorrected MAE**	**BBC MAE**	**Imp. (%)**
A	0.0	0.2625	0.1569	40.2	0.1984	0.1269	36.1
	0.5	0.3065	0.1692	44.8	0.2355	0.1391	40.9
	0.8	0.2827	0.1557	44.9	0.2349	0.1401	40.3
B	0.0	0.4931	0.3145	36.2	0.3891	0.2348	39.6
	0.5	0.5031	0.3045	39.5	0.4027	0.2478	38.5
	0.8	0.4388	0.2623	40.2	0.3411	0.2180	36.1
C	0.0	0.0474	0.0569	−20.1	0.0321	0.0337	−4.7
	0.5	0.0546	0.0641	−17.5	0.0389	0.0392	−0.8
	0.8	0.0543	0.0646	−19.0	0.0441	0.0472	−7.2

Note. Each setting includes $R = 30$ independently fitted trees. Uncorrected MAE and BBC MAE refer to terminal-node response standard deviations. Imp. denotes percentage MAE reduction following BBC. Negative values indicate increased estimation error.

Table A8. Sensitivity of classification terminal-node probability estimation to predictor dependence.

		$n = 500$			$n = 1000$		
Model	ρ	**Uncorrected MAE**	**BBC MAE**	**Imp. (%)**	**Uncorrected MAE**	**BBC MAE**	**Imp. (%)**
A	0.0	0.0852	0.0708	16.9	0.0611	0.0496	18.9
	0.5	0.0728	0.0607	16.6	0.0810	0.0661	18.5
	0.8	0.0524	0.0455	13.3	0.0624	0.0499	20.1
B	0.0	0.0822	0.0639	22.3	0.0634	0.0523	17.6
	0.5	0.0781	0.0664	15.0	0.0725	0.0608	16.1
	0.8	0.0637	0.0532	16.4	0.0530	0.0452	14.6
C	0.0	0.0301	0.0282	6.5	0.0179	0.0174	3.2
	0.5	0.0316	0.0298	5.5	0.0212	0.0194	8.6
	0.8	0.0331	0.0340	−2.6	0.0237	0.0232	2.3

Note. Each setting includes $R = 30$ independently fitted trees. Uncorrected MAE and BBC MAE refer to terminal-node event probabilities. Imp. denotes percentage MAE reduction. Negative values indicate deterioration.

Table A8 shows that BBC reduced probability MAE for classification Models A and B across all examined dependence levels. The improvements ranged from 13.3% to 22.3%.

For classification Model C, the improvements were smaller and less consistent. At $\rho = 0.8$ and $n = 500$, BBC increased probability MAE by approximately 2.6%.

Overall, these findings suggest that the benefits of BBC for nonlinear Models A and B persist under the Gaussian-copula predictor-dependence structures examined. However, BBC does not uniformly improve estimation, particularly in Model C, where the uncorrected error is relatively small. The experiment provides targeted evidence of robustness to predictor dependence rather than a guarantee under arbitrary correlation structures or higher-dimensional settings.

Appendix F. Illustration of Observation-Overlap Mapping

This appendix illustrates how bootstrap terminal-node discrepancies are mapped to an original CART terminal node using observation overlap, as defined in Equation (6).

Consider an original terminal node t containing 10 observations, with index set $\mathcal{I}_t = \{1, 2, \ldots, 10\}$. After routing these original observations through a fitted bootstrap tree, suppose that five belong to bootstrap terminal node t'_1, three to t'_2, and two to t'_3. The resulting overlap weights are 0.5, 0.3, and 0.2, respectively.

Original CART terminal node t

$$\boxed{\mathcal{I}_t = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}}$$

$$\downarrow$$

Route the original observations through bootstrap tree T_b

Bootstrap node	Shared observation indices	Overlap weight
t'_1	$\{1, 2, 3, 4, 5\}$	$5/10 = 0.5$
t'_2	$\{6, 7, 8\}$	$3/10 = 0.3$
t'_3	$\{9, 10\}$	$2/10 = 0.2$

$$\downarrow$$

Weighted bootstrap-node correction

$$c_t^{(b)} = 0.5\delta_{bt'_1} + 0.3\delta_{bt'_2} + 0.2\delta_{bt'_3}.$$

$$\widehat{c}_t = \frac{1}{B}\sum_{b=1}^{B} c_t^{(b)}.$$

Figure A1. Schematic illustration of observation-overlap mapping between an original CART terminal node and terminal nodes of a fitted bootstrap tree. The original observations are routed through the bootstrap tree, and each bootstrap-node discrepancy contributes to the original-node correction according to the proportion of shared observations. The counts are illustrative.

For this bootstrap replication, the overlap weights sum to one: $0.5 + 0.3 + 0.2 = 1$. The same mapping is performed across bootstrap replications, and the weighted discrepancies are averaged to obtain the correction for the original terminal node.

Appendix G. Directional Assessment of Estimation Errors

The MAE and RMSE measures reported in the main simulation study quantify the magnitude of estimation errors but do not distinguish systematic overestimation from underestimation. To complement these measures, we examine signed estimation discrepancies relative to the independent test-sample estimates.

For a terminal-node estimate $\hat{\theta}_j$ and its independent test-sample reference $\hat{\theta}_j^{\text{test}}$, define the signed discrepancy as

$$d_j = \hat{\theta}_j - \hat{\theta}_j^{\text{test}}.$$

Positive values indicate overestimation relative to the reference, whereas negative values indicate underestimation. The mean signed discrepancy, $\bar{d} = J^{-1} \sum_{j=1}^{J} d_j$, provides a directional summary across the J terminal-node records in a simulation setting. These discrepancies are evaluated separately for uncorrected and BBC-corrected estimates.

Figure A2 shows that uncorrected regression standard deviation estimates tend to underestimate the independent test-sample reference in Models A and B. BBC generally shifts these discrepancies toward zero, although positive discrepancies may arise following correction. For Model C, the uncorrected discrepancies are comparatively small, and BBC does not consistently improve agreement.

Figure A3 shows that classification proportion discrepancies are generally centered near zero, with narrower distributions at larger sample sizes. The directional effects of BBC are less pronounced and vary across model settings.

Figures A4 and A5 provide complementary visual assessments of the signed discrepancies. Together with the signed-error distributions, these comparisons illustrate the direction of BBC adjustments and identify settings in which the correction may overshoot the independent test-sample reference. The results should be interpreted alongside the MAE and RMSE summaries rather than as replacements for those measures.

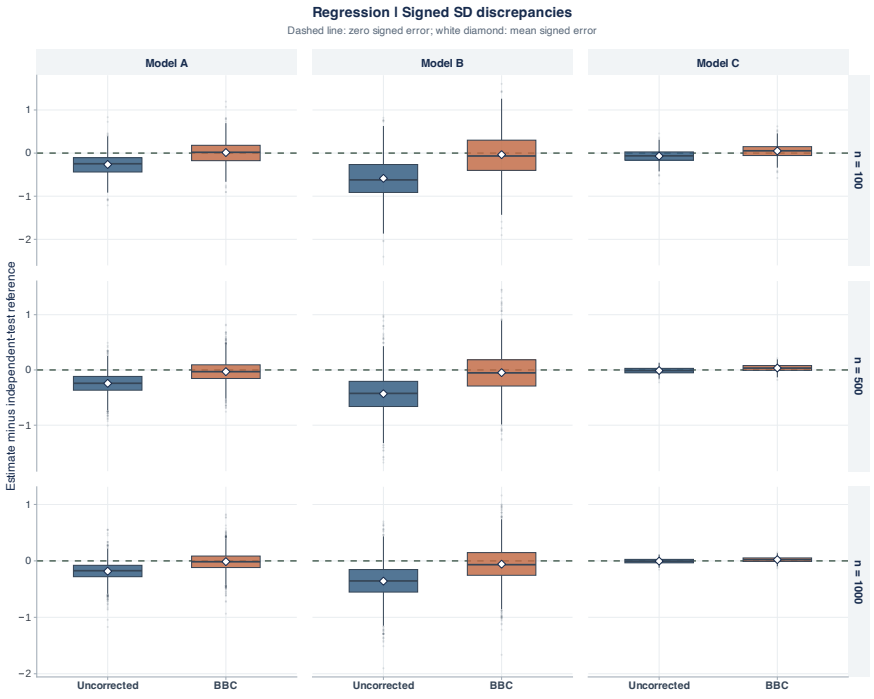


Figure A2. Signed discrepancies in terminal-node standard deviation estimates for regression Models A–C at sample sizes $n = 100$, 500, and 1000. Discrepancies are calculated relative to independent test-sample estimates. The dashed horizontal line indicates zero discrepancy, and white diamonds represent mean signed discrepancies.

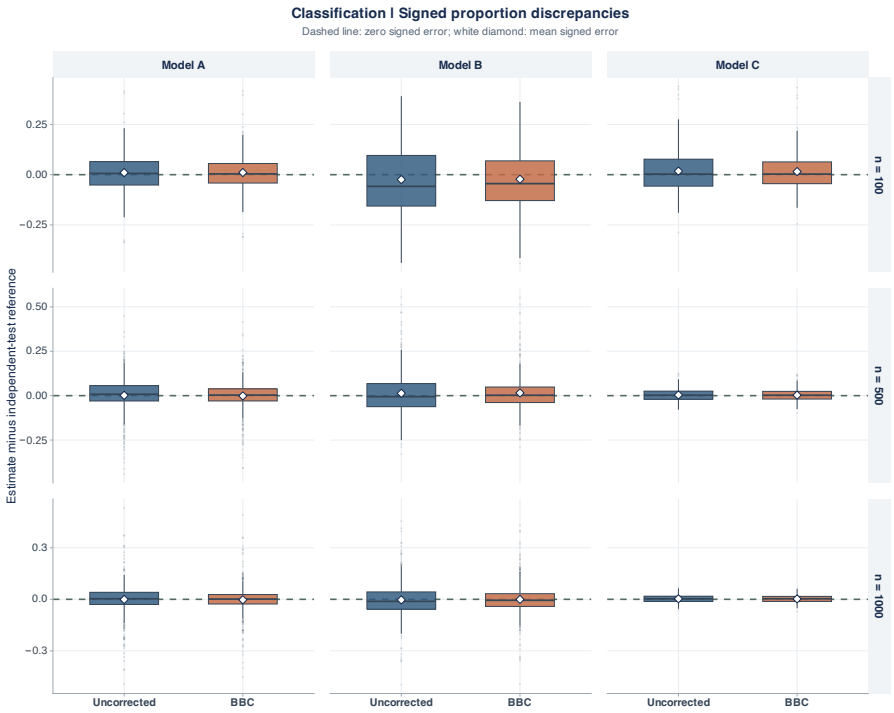


Figure A3. Signed discrepancies in terminal-node class proportion estimates for classification Models A–C at sample sizes $n = 100$, 500, and 1000. The dashed horizontal line indicates zero discrepancy, and white diamonds represent mean signed discrepancies.

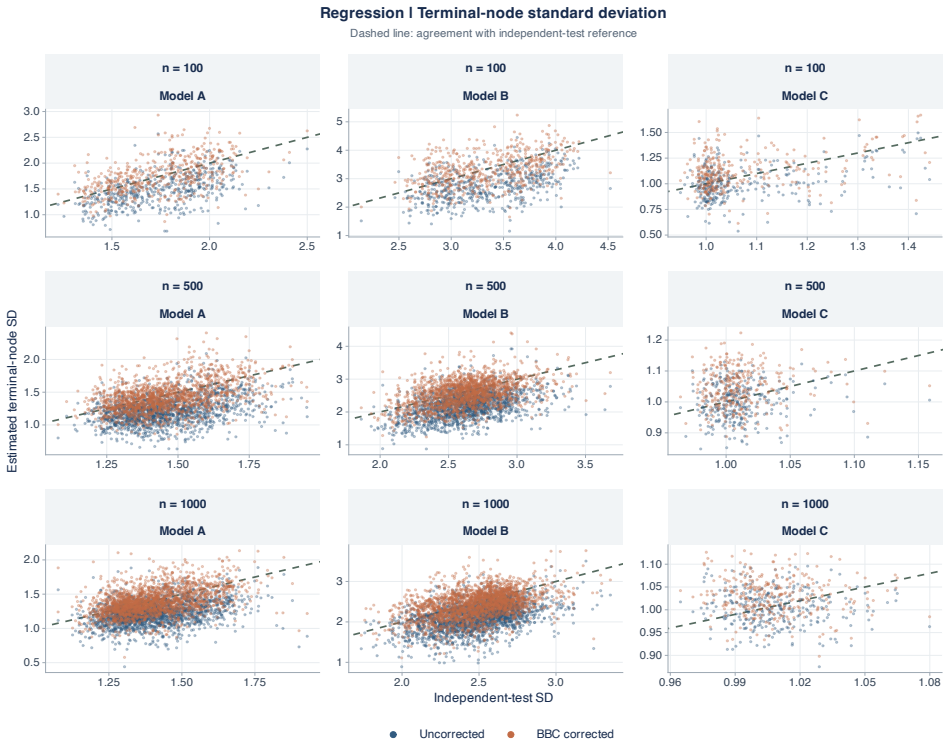


Figure A4. Comparison of uncorrected and BBC-corrected terminal-node standard deviation estimates with independent test-sample estimates for regression Models A–C. The dashed line represents perfect agreement. Points above and below this line indicate positive and negative signed discrepancies, respectively. Axis scales vary across panels.

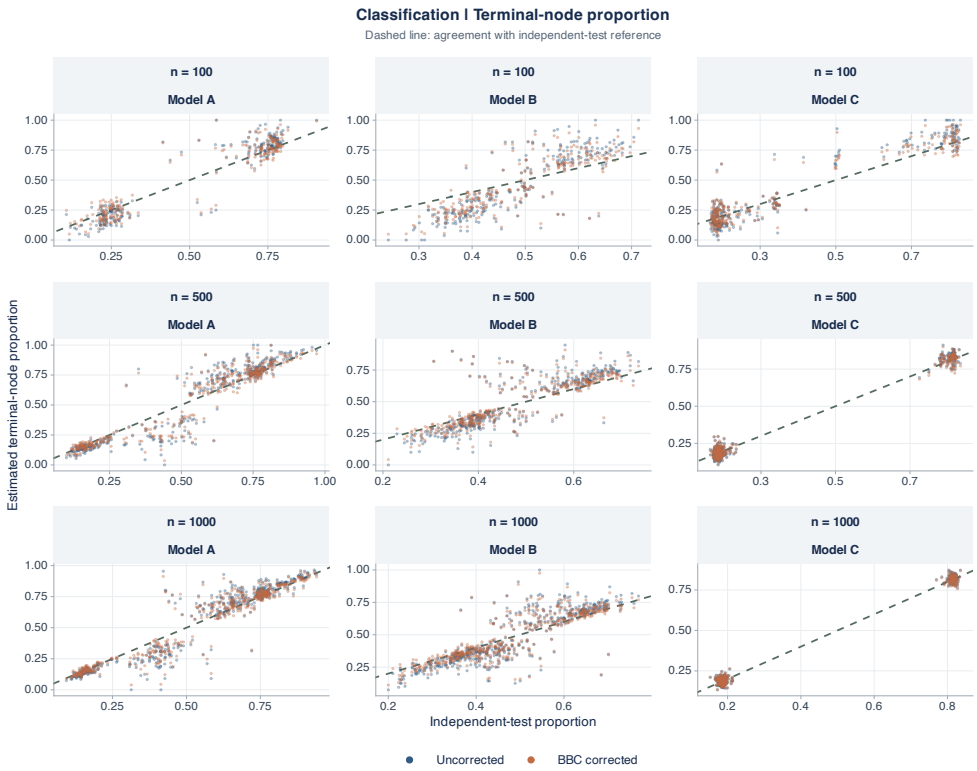


Figure A5. Comparison of uncorrected and BBC-corrected terminal-node class proportion estimates with independent test-sample estimates for classification Models A–C. The dashed line represents perfect agreement. Points above and below this line indicate overestimation and underestimation relative to the independent test-sample reference. Axis scales vary across panels.

Appendix H. Empirical Confidence-Interval Coverage and Width

This appendix reports the detailed numerical results underlying the confidence-interval coverage analysis in Section 4.6. Coverage denotes the empirical proportion of terminal-node intervals containing the corresponding independent-reference estimate, averaged first within each fitted tree and then across Monte Carlo replications. Width denotes the average confidence-interval length calculated using the same aggregation procedure. Coverage gains are expressed in percentage points, whereas width increases are expressed as percentages relative to the uncorrected intervals.

Table A9. Empirical coverage and average width of nominal 95% confidence intervals in regression-tree simulations.

		Empirical Coverage			Average Interval Width		
n	**Model**	**Uncorr.**	**BBC**	**Gain (pp)**	**Uncorr.**	**BBC**	**Increase (%)**
100	A	0.856	0.901	4.55	1.340	1.590	18.6
	B	0.816	0.894	7.87	2.530	3.040	19.9
	C	0.894	0.921	2.67	0.709	0.795	12.1
500	A	0.803	0.872	6.85	0.922	1.090	18.0
	B	0.834	0.894	6.01	1.710	2.020	17.8
	C	0.923	0.937	1.33	0.319	0.333	4.60

Table A9. *Cont.*

		Empirical Coverage			Average Interval Width		
n	Model	Uncorr.	BBC	Gain (pp)	Uncorr.	BBC	Increase (%)
1000	A	0.813	0.874	6.08	0.771	0.884	14.6
	B	0.817	0.874	5.69	1.430	1.640	14.8
	C	0.910	0.923	1.33	0.226	0.232	2.66

Note. Uncorr. denotes conventional uncorrected CART confidence intervals; BBC denotes intervals constructed using bootstrap bias-corrected terminal-node standard deviations. Gain (pp) is the difference in empirical coverage expressed in percentage points. Increase (%) is the relative increase in average interval width. Results use $R = 100$ Monte Carlo replications, $B = 50$ bootstrap resamples, and independent reference samples of size $n_0 = 10{,}000$. Displayed values are rounded independently.

Table A10. Empirical coverage and average width of nominal 95% confidence intervals in classification-tree simulations.

		Empirical Coverage			Average Interval Width		
n	Model	Uncorr.	BBC	Gain (pp)	Uncorr.	BBC	Increase (%)
100	A	0.781	0.852	7.17	0.229	0.241	5.19
	B	0.601	0.734	13.4	0.266	0.279	5.16
	C	0.795	0.889	9.42	0.244	0.260	6.87
500	A	0.769	0.835	6.65	0.166	0.173	4.36
	B	0.683	0.830	14.7	0.166	0.170	2.21
	C	0.917	0.947	3.00	0.123	0.124	1.15
1000	A	0.770	0.862	9.15	0.145	0.149	3.21
	B	0.682	0.835	15.3	0.151	0.154	1.88
	C	0.943	0.943	0.00	0.0870	0.0875	0.567

Note. Uncorr. denotes conventional CART Wald-type confidence intervals; BBC denotes intervals using the corrected class probabilities and their associated Bernoulli variances. Interval endpoints are restricted to $[0, 1]$. Gain (pp) is the difference in empirical coverage expressed in percentage points. Increase (%) is the relative increase in average interval width. Results use $R = 100$ Monte Carlo replications, $B = 50$ bootstrap resamples, and independent reference samples of size $n_0 = 10{,}000$. Displayed values are rounded independently.

Appendix I. Robustness to Heavy-Tailed Regression Errors

To assess the robustness of bootstrap bias correction (BBC) to departures from the Gaussian error assumption, we extended the regression simulation study to two non-normal error distributions: a Student's t distribution with three degrees of freedom and a right-skewed lognormal distribution. Both distributions were standardized to have mean zero and variance one, allowing the effect of error distribution to be examined without changing the overall error variance.

Specifically, the standardized Student's t errors were generated as

$$\varepsilon = \frac{T_3}{\sqrt{3}}, \qquad T_3 \sim t_3. \tag{A7}$$

The standardized lognormal errors were generated as

$$\varepsilon = \frac{\exp(Z) - \exp(1/2)}{\sqrt{(e-1)e}}, \qquad Z \sim N(0,1). \tag{A8}$$

The Student's t distribution represents symmetric heavy-tailed errors, whereas the standardized lognormal distribution introduces right-skewness and heavy upper-tail behavior.

All other aspects of the original simulation design were retained. Specifically, we considered regression Models A–C with training sample sizes $n \in \{100, 500, 1000\}$. Each setting included $R = 100$ Monte Carlo replications, and the BBC procedure used $B = 50$

bootstrap resamples. Independent reference samples of size $n_0 = 10{,}000$ were used to approximate the population standard deviations within the selected terminal regions.

Performance was evaluated using mean absolute error (MAE) and root mean squared error (RMSE) of the uncorrected and BBC-adjusted terminal-node standard deviations relative to the corresponding reference-sample estimates. Consistent with the original simulation analysis, discrepancies were pooled across terminal nodes and Monte Carlo replications.

Table A11 summarizes the performance of BBC under the two non-normal regression error distributions. For Models A and B, BBC consistently reduced both MAE and RMSE across all training sample sizes and error distributions. The reductions in MAE ranged from approximately 18.4% to 40.6%, indicating that the correction retained substantial benefits under the examined symmetric heavy-tailed and right-skewed error mechanisms.

Table A11. Performance of uncorrected CART and BBC-adjusted terminal-node standard-deviation estimates under heavy-tailed regression errors.

Error	n	Model	MAE:CART	MAE:BBC	RMSE:CART	RMSE:BBC	MAE Gain (%)
Student t_3	100	A	0.369	0.286	0.472	0.442	22.4
		B	0.700	0.440	0.809	0.558	37.2
		C	0.239	0.203	0.288	0.266	14.8
	500	A	0.329	0.235	0.481	0.422	28.5
		B	0.492	0.316	0.586	0.410	35.8
		C	0.148	0.157	0.213	0.241	−5.9
	1000	A	0.262	0.202	0.343	0.314	23.1
		B	0.408	0.264	0.531	0.426	35.2
		C	0.138	0.145	0.191	0.217	−4.9
Lognormal	100	A	0.402	0.292	0.468	0.391	27.4
		B	0.723	0.430	0.837	0.551	40.6
		C	0.306	0.283	0.364	0.353	7.4
	500	A	0.346	0.258	0.423	0.372	25.3
		B	0.517	0.342	0.621	0.475	34.0
		C	0.214	0.227	0.271	0.311	−5.8
	1000	A	0.303	0.247	0.414	0.411	18.4
		B	0.426	0.286	0.518	0.412	32.8
		C	0.166	0.187	0.226	0.273	−12.7

Note. CART denotes the conventional uncorrected terminal-node standard deviation, whereas BBC denotes the bootstrap bias-corrected estimate. The percentage improvement in MAE is calculated as $100 \times (\mathrm{MAE}_{\mathrm{CART}} - \mathrm{MAE}_{\mathrm{BBC}})/\mathrm{MAE}_{\mathrm{CART}}$. Positive values indicate improved accuracy under BBC, whereas negative values indicate worse performance. Results are based on 100 Monte Carlo replications per setting and 50 bootstrap resamples. Numerical values are rounded.

For the simpler tree-structured Model C, the results were more mixed. At $n = 100$, BBC improved both MAE and RMSE under both error distributions. However, at $n = 500$ and $n = 1000$, the uncorrected estimates performed better, with BBC increasing both MAE and RMSE. The deterioration was particularly pronounced under lognormal errors at $n = 1000$, where the MAE increased by approximately 12.7% relative to the uncorrected estimator.

These findings suggest that BBC can provide meaningful improvements in terminal-node standard-deviation estimation under departures from Gaussian errors, particularly for nonlinear regression mechanisms where adaptive tree construction introduces appreciable estimation distortion. Nevertheless, the correction does not uniformly improve estimation accuracy. In settings where the underlying regression structure is well represented by a simple tree, the additional correction may be unnecessary or counterproductive.

Overall, the heavy-tailed experiments support the usefulness of BBC beyond the Gaussian simulation setting while highlighting its dependence on the underlying regression structure and sample size.

References

1. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
2. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
3. Lakshminarayanan, B. Decision Trees and Forests: A Probabilistic Perspective. Ph.D. Thesis, UCL (University College London), London, UK, 2016.
4. Efron, B. Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* **2014**, *109*, 991–1007. [CrossRef] [PubMed]
5. Kuchibhotla, A.K.; Kolassa, J.E.; Kuffner, T.A. Post-Selection Inference. *Annu. Rev. Stat. Its Appl.* **2022**, *9*, 505–527. [CrossRef]
6. Loh, W.Y.; Man, M.; Wang, S. Subgroups from Regression Trees with Adjustment for Prognostic Effects and Post-Selection Inference. *Stat. Med.* **2019**, *38*, 545–557. [CrossRef] [PubMed]
7. Neufeld, A.C.; Gao, L.L.; Witten, D.M. Tree-Values: Selective Inference for Regression Trees. *J. Mach. Learn. Res.* **2022**, *23*, 1–43.
8. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
9. Loh, W.Y. Calibrating confidence coefficients. *J. Am. Stat. Assoc.* **1987**, *82*, 155–162. [CrossRef]
10. Loh, W.Y. Bootstrap calibration for confidence interval construction and selection. *Stat. Sin.* **1991**, *1*, 477–491.
11. LeBlanc, M.; Crowley, J. Survival Trees by Goodness of Split. *J. Am. Stat. Assoc.* **1993**, *88*, 457–467. [CrossRef]
12. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
13. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002. [CrossRef]
14. Quaye, G.E. Making Valid Inferences with Decision Tree. Master's Thesis, The University of Texas at El Paso, El Paso, TX, USA, 2021.
15. Su, X.; Zhou, T.; Yan, X.; Fan, J.; Yang, S. Interaction trees with censored survival data. *Int. J. Biostat.* **2008**, *4*, 2. [CrossRef] [PubMed]
16. Su, X.; Tsai, C.L.; Wang, H.; Nickerson, D.M.; Li, B. Subgroup Analysis via Recursive Partitioning. *J. Mach. Learn. Res.* **2009**, *10*, 141–158.
17. Zhou, R.; Xie, K.; Liu, L.; Xu, Z.; Ding, J.; Su, X. Enhanced Survival Trees. *arXiv* **2025**, arXiv:2509.18494.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.