When Causal Forests Mislead

Evaluating the precision of Confidence Intervals

Rares Iordan[1]
Supervisor(s): Jesse Krijthe[1]**, Rickard Karlsson**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Rares Iordan
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Rickard Karlsson, Ricardo Marroquim

Abstract

This study tackles an important issue in evaluating the reliability of confidence intervals in causal forests by examining how data characteristics and hyperparameters influence actual coverage rates compared to theoretical benchmarks. Using synthetic data sets with polynomial treatment effects, Sobol sampling, High-Dimensional Model Representation (HDMR), and comprehensive grid searches, the study assesses causal forest performance in different data contexts.

A primary discovery is the identification of a practical limit for reliable confidence interval coverage: When the sum of confounders and effect modifiers exceeds 4, coverage rates drop considerably below 80%, even for simple treatment effect functions. This limitation remains steady despite substantial increases in computational resources.

The examination of hyperparameters revealed that the most influential parameters are the maximum tree depth and the balance tolerance in splits, which demonstrate substantial changes in performance, both of which performed best at their maximums (unlimited and 0.5, respectively). Other key suggestions involve increasing the training data fraction per tree from 0.45 to 0.5, keeping the minimum impurity decrease threshold at 0.0, and utilizing at least ≈ 2400 trees to meet theoretical expectations.

In addition, this paper did not identify any noteworthy interaction between tree count and sample size. As a result, both of these characteristics can be optimized independently of each other.

These findings provide systematic guidelines for practitioners to assess when causal forest confidence intervals are reliable and how to optimize them, bridging the gap between theoretical guarantees and practical performance.

1 Introduction

Machine learning algorithms usually aim to predict output or outcomes based on given inputs. However, in many fields, such as healthcare, predicting outcomes is not enough, as it is often necessary to estimate the causal effects of particular interventions, also called treatments (e.g., what would change if a treatment were to be administered to a patient or not). Simply predicting outcomes rather than understanding causality can lead to harmful decisions (e.g., a model could predict poor outcomes for patients without recognizing that those patients might already have been in a worse-off position, leading to inappropriate withholding of treatments).

These causal effects have typically been measured using the average treatment effect (ATE) of a given intervention, which only measures the expected difference in outcomes between treated and untreated populations throughout the sample. However, there is growing interest in predicting not only ATE, but also the conditional average treatment effect (CATE), which is how treatments affect different subpopulations differently. Understanding heterogeneous effects would greatly aid healthcare workers, allowing for personalized predictions, such as which patients would benefit the most from a given treatment. Without this understanding, there is a risk of applying treatments to subgroups of people for whom the treatment is ineffective or harmful.

When dealing with observational data, researchers have to deal with various issues, such as confounding variables that affect both the treatment assignment and the outcome (e.g., more seriously ill people are more likely to receive treatment and less likely to recover) [1]. This problem breaks the independence between treatment assignments and outcomes, allowing machine learning algorithms to identify associations instead of causations accidentally. This is why algorithms that can account for such challenges are needed.

Random forests, introduced by Breiman [2], are a machine learning technique widely used for prediction due to their flexibility and capacity to manage complex nonlinear interactions between variables. Building upon this framework, Wager and Athey proposed Causal Forests [3], a method for heterogeneous treatment effect estimation that represents the first variant of random forests to provide proven asymptotically valid confidence intervals. Building on the work by Wager and Athey [3], Athey *et al.* [4] advanced the field by creating Generalized Random Forests (GRF), a framework that not only generalizes causal forests through local moment conditions but also offers a more computationally efficient implementation, solidifying GRF causal forests as the standard for causal inference with random forests while preserving theoretical guarantees of valid confidence intervals.

However, despite the theoretical foundations and practical improvements, a significant gap in the literature became apparent. There are few studies exploring the behavior of confidence intervals in causal forests when exposed to different data characteristics or how different hyperparameters affect the coverage intervals. This is in part due to the fact that the main body of research is focused on prediction accuracy or optimizing hyperparameters for prediction accuracy rather than uncertainty quantification, such as the study by Saito and Yasui [5]. Systematic evaluations of the model showed that the actual coverage rate of the confidence intervals is significantly lower than expected, despite theoretical guarantees [6].

This discrepancy is alarming because, for practical applications, it is essential for practitioners to know how to adjust the hyperparameters to optimize for coverage rates and discern when the confidence intervals can be trusted. Moreover, although existing studies examine performance across different levels of noise, sample size, and dimensionality, a more thorough investigation into data characteristics, such as the relationship between the specific numbers of confounders, effect modifiers, instruments, or the type of treatment effect function, has not yet been conducted.

The relationship between data characteristics, sample size, and number of trees is also poorly understood. The theoretical framework suggests that, for the confidence intervals to hold, a sufficient number of trees is necessary to make the Monte Carlo noise negligible [4], [7]. However, it is unclear how this requirement interacts with other factors, such as the size of the training dataset.

This study aims to fill these gaps by providing a comprehensive empirical analysis of the influence of various data characteristics on coverage rates in causal forests. It also assesses the effect of hyperparameters on coverage rather than point-estimate accuracy, addressing the following question:

What factors affect the actual coverage rates of the confidence interval estimation in causal forests, and what are the optimal configuration parameters for different data scenarios when limited to polynomial treatment effect functions of low or medium order (1–5), low to high confounding strength (1–10), and low to medium dimensionality (1–21)?

1. How sensitive is the confidence interval coverage rate of causal forests to the number of confounders, instruments, effect modifiers, polynomial treatment effect complexity, and confounding strength? When does achieving 95% coverage become unfeasible?
2. Which hyperparameters individually influence the coverage rate of the confidence interval and in what manner do they exert this influence?
3. How do tree count and sample size interact to affect confidence interval coverage rates in causal forests?

This study makes a contribution by pinpointing a practical limit at which the reliability of coverage rates experiences a substantial decrease (at 4 combined confounders/effect modifiers), rendering it impractical to achieve the expected values. It also offers empirical recommendations for hyperparameters concerning confidence intervals instead of focusing on prediction accuracy, and revealed no significant interactions between the number of trees and the amount of training data.

2 Background

In order to address the research question at hand, it is necessary to understand the theoretical aspects of the estimation of the heterogeneous treatment effect, the causal forests, and the various methods used in this study.

2.1 Heterogeneous Treatment Effect Estimation

2.1.1 Formalization

Causal inference can be formulated using the potential outcomes framework [8]. For each unit i, potential outcomes are described as $Y_i^{(0)}$ and $Y_i^{(1)}$, indicating observations under control and treatment conditions, respectively. The treatment effect for an individual is given by $\tau_i = Y_i^{(1)} -$

$Y_i^{(0)}$, although only $Y_i = W_i Y_i^{(1)} + (1 - W_i) Y_i^{(0)}$ can be observed, where $W_i \in \{0, 1\}$ denotes whether the treatment was applied to the unit i.

The Conditional Average Treatment Effect (CATE) is defined as:

$$\tau(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} | X = x] \tag{1}$$

CATE encapsulates the variation of treatment effects throughout the covariate space X, thus enabling heterogeneous causal inference.

2.1.2 Necessary assumptions for Observational Data

To draw valid causal inferences from observational data and estimate the CATE, researchers must rely on key assumptions to compensate for the lack of randomization. The following three assumptions are fundamental for identifying causal effects [9]:

1. **Unconfoundedness**: The assignment of treatment is independent of the potential outcomes given the observed covariates. This assumption ensures that there are no unmeasured confounders, factors that influence both treatment assignment and outcomes. In the absence of unconfoundedness, any estimated treatment effect may be biased due to omitted variable bias, which is particularly important in the context of CATE estimation, since any violations could mix genuine treatment heterogeneity with omitted variable bias.

2. **Positivity**: Every individual has a positive probability of receiving the treatment or not. This assumption is crucial to ensure that comparisons can actually be made across treatment groups for all subpopulations. In the absence of positivity, estimating causal effects for individuals in covariate regions where only one treatment is observed becomes impossible, which directly affects CATE estimation by creating regions where $\tau(x)$ cannot be reliably estimated due to insufficient overlap between treatment groups.

3. **Stable Unit Treatment Values Assumption (SUTVA)**:

   (a) The assignment of treatment to units does not interfere with the outcome of other units. This prevents spillover effects that would violate the independence of units.

   (b) There are no hidden variations of treatment. This rules out multiple, unaccounted-for versions of the same treatment that could yield different effects.

   SUTVA is essential to ensure that the treatment effect is well defined and solely attributable to the treatment assigned, and, for heterogeneous treatment effects, it ensures that $\tau(x)$ depends only on individual characteristics and not on interference or hidden treatment variations.

Similarly to all techniques for estimating CATE from observational data, Causal Forests also rely on these assumptions being valid.

2.2 Causal Forests

Causal forests, as introduced by Wager and Athey [3], build on random forests [2] by using "honest" trees to estimate heterogeneous treatment effects. An "honest" tree uses the outcome variable Y_i either for splitting or leaf estimation, but not both, thus allowing valid statistical inference. A method for constructing "honest" trees is employing double sample trees [3], where the training data is split into two halves, one used for structure and one used for estimation. The procedure can be explained as follows:

Procedure: Double-Sample Trees for CATE Estimation [3]

1. **Subsampling and Partitioning:**
   - Draw a random subsample $S \subset \{1, \ldots, n\}$ of size s without replacement, where n is the size of the training dataset.
   - Partition S into disjoint sets I and J where $|I| = \lfloor s/2 \rfloor$ and $|J| = \lceil s/2 \rceil$
2. **Tree Structure Construction (using J-sample):**
   - For splitting, use features X_i and treatments W_i from both I and J samples, and Y_i from J-sample only
   - Select splits that maximize variance of $\hat{\tau}(X_i)$ for $i \in J$
   - Continue while valid splits are possible (i.e., until splitting would create leaves with $< k$ observations from either treatment group)
3. **Leaf Effect Estimation (using I-sample only):**
   - For leaf L containing test point x, the estimated CATE is the difference in mean outcomes between treated and control units.

The Causal Forest analyzed in this paper is based on a more sophisticated iteration of the Causal Forests developed by Wager and Athey [3], known as General Random Forests (GRF) [4]. While the original causal forests are specifically designed for estimating treatment effects, the GRF adopts a broader mathematical approach known as "local moment conditions," allowing it to address a wider range of problems, not just causal inference.

However, the aforementioned procedure is "almost equivalent to a generalized random forest [...], the only substantive differences being that they split using the exact loss criterion rather than [the] gradient-based loss criterion, and let each tree compute its own treatment effect estimate rather than using the [adaptive] weighting scheme" [4], where *gradient-based loss criterion* serves as a computational optimization technique to estimate the splitting criterion, and *adaptive weighting scheme* aims to reduce the bias stemming from employing local moment conditions instead of direct CATE estimation. Although these implementation specifics impact computational performance and might affect point estimates, they do not fundamentally change the core method or the confidence interval construction, which are the main focus of this investigation and thus will not be elaborated on in this section.

2.2.1 Confidence Intervals

Generalized Random Forests utilize the bootstrap of little bags [4] to compute confidence intervals by providing variance estimations for predictions. This methodology bypasses the complexities of conventional bootstrap methods that would require the regeneration of the entire forest by taking advantage of the subsampling that is naturally part of the forest construction process. The core insight is that distinct trees, each trained on different sub-samples, inherently reveal insight into how estimates might fluctuate with different training datasets.

A naive approach might be to calculate the variance of single-tree predictions. However, this would mix two separate sources of variation: (1) true sampling uncertainty, showing how predictions would shift with different datasets, and (2) Monte Carlo noise due to the finite number of trees used.

The bootstrap of little bags strategy addresses these issues by grouping trees and applying a variance decomposition that distinguishes between-group variation, indicative of sampling uncertainty, from within-group variation, which is characteristic of Monte Carlo noise [10].

The variance estimation procedure works as follows:

1. **Little Bag Construction:** Partition the B trees into groups of size ℓ, creating $G = B/\ell$ groups. Each group represents an independent "experiment" using different half-samples of the training data. By employing half-samples, an ideal equilibrium between sample size and independence is accomplished, which is essential for estimating variance.

2. **Variance Decomposition:** The variance is estimated by measuring how much between-group estimates vary, while correcting for the finite number of trees within each group by subtracting a scaled version of the within-group variance, thus isolating the true sampling uncertainty:

$$\hat{V}_{BLB}(x) = \underbrace{\frac{1}{G}\sum_{g=1}^{G}(\hat{\tau}^g(x) - \hat{\tau}(x))^2}_{\text{Between-group variance}} - \underbrace{\frac{1}{\ell - 1} \cdot \frac{1}{G}\sum_{g=1}^{G}\frac{1}{\ell}\sum_{b \in g}(\hat{\tau}_b(x) - \hat{\tau}^g(x))^2}_{\text{Within-group correction}} \tag{2}$$

Here, $\hat{\tau}^g(x) = \frac{1}{\ell}\sum_{b \in g}\hat{\tau}_b(x)$ is the average treatment effect estimate within group g, and $\hat{\tau}_b(x)$ is the estimate from individual tree b.

In addition, under the assumptions specified by Athey *et al.* [4], the treatment effect estimates are asymptotically Gaussian and unbiased [4], thus enabling the construction of valid confidence intervals (95%).

$$\hat{\tau}(x) \pm 1.96\sqrt{\hat{V}_{BLB}(x)} \tag{3}$$

2.3 Sobol' Sampling

To systematically investigate how causal forest confidence interval coverage rates are affected by various data characteristics and their interactions, it is necessary to efficiently explore a multidimensional parameter space. This exploration presents a sampling challenge: with multiple parameters each taking various values, a comprehensive grid search becomes computationally prohibitive, while pure random sampling may leave important regions of the parameter space inadequately covered due to clustering and gaps.

Sobol sequences, first introduced by Sobol' in 1967 [11], address this challenge through deterministic quasirandom sampling designed to systematically cover multidimensional spaces. These sequences demonstrate a more uniform distribution of sample points across the parameter space compared to random sampling [12], [13]. The method works by following a predetermined pattern that places each new sample point in the location that best fills the gaps left by previous points.

In addition, empirical comparisons have demonstrated the superiority of Sobol sequences over alternative sampling methods. Tarantola et al. [14] conducted systematic comparisons across multiple test functions and found that "in almost all cases investigated here, the Sobol' design performs better" than Latin Hypercube Sampling — another widely used sampling technique, which divides each parameter dimension into equal intervals and ensures one sample per interval — with the results indicating that "the Sobol' design was consistently superior." Similarly, Sudret et al. [15] showed that Sobol sequences "performed globally better in all the numerical experiments" compared to Monte Carlo and Latin Hypercube sampling approaches.

Furthermore, Sobol sequences provide reduced variability in parameter space coverage between different sampling runs, while being significantly more reproducible than Latin Hypercube Sampling [16].

For these reasons, Sobol sampling was utilized to efficiently traverse the vast multidimensional parameter space.

2.4 High-Dimensional Model Representation (HDMR)

To address the research question on the sensitivity of coverage rates of causal forest confidence intervals to different data characteristics, it is essential to perform a sensitivity analysis. This analysis helps to assess how individual parameters and their interactions impact the variability of model performance. However, using standard methods such as Sobol analysis presents substantial computational challenges, as these traditional approaches generally require a large number of model evaluations, which are not feasible due to the resource-demanding characteristics of causal forest experiments. This is why an alternative was needed that can perform accurate sensitivity analyses with fewer data points.

High-Dimensional Model Representation (HDMR) [17] addresses these limitations and enables efficient sensitivity analysis through a metamodeling framework using hierarchical decomposition. This metamodeling approach means HDMR builds a simplified mathematical approximation (surrogate model) of the complex causal forest behavior, allowing the analysis of sensitivity without running the full expensive simulations repeatedly. Rather than directly sampling the parameter space extensively, HDMR constructs this surrogate model using systematic expansion:

$$f(x_1, x_2, \ldots, x_d) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \sum_{i<j<k} f_{ijk}(x_i, x_j, x_k) + \ldots \quad (4)$$

where f_0 represents the mean output, $f_i(x_i)$ captures the independent effect of parameter x_i, $f_{ij}(x_i, x_j)$ represents interactions between parameters, and higher order terms account for complex parameter interactions [18].

An important setting for this approach is `max_order`, which restricts the decomposition to the specified order. Higher values detect higher-order interactions; however, it is worth noting that increasing this value also requires more samples for valid results.

The method returns sensitivity indices that represent what percentage of the variance can be explained by the parameter in question.

Compared to classical approaches like Sobol indices, HDMR is often more efficient and requires fewer samples [18]. In addition, according to SALib's documentation, a sensitivity analysis library in Python, "HDMR becomes extremely useful when the computational cost of obtaining sufficient Monte Carlo samples are prohibitive, as may be the case with Sobol's method." [19]

The SALib implementation incorporates bootstrap resampling to provide confidence intervals for sensitivity indices, enabling robust uncertainty quantification [19].

Thus, SALib's implementation of HDMR was selected because of its computational efficiency and provision of confidence intervals for identifying the most influential parameters affecting causal forest performance.

3 Methodology

3.1 Data Generating Process (DGP)

This research employs a polynomial data-generating process provided by the supervisor that generates synthetic observational datasets with known causal structure. The DGP creates data where the true treatment effect function follows a polynomial form of controllable complexity, allowing systematic evaluation of causal forest performance under varying data characteristics including the following parameters pertinent to the paper:

- **Polynomial degree** – Controls treatment effect complexity
- **Confounding strength** – Determines the magnitude of confounding bias
- **Number of confounders** – Variables affecting both treatment and outcome
- **Number of instruments** – Variables affecting only treatment assignment
- **Number of effect modifiers** – Variables interacting with the treatment effect

3.2 Exhaustive Grid Search

To comprehensively examine the parameter space and address certain aspects of the questions, an exhaustive grid search methodology was employed when needed. This involved creating all conceivable combinations of the defined parameters to assess causal forest performance based on three metrics: mean squared error, confidence interval coverage rate, and width. Each experimental setup was executed multiple times with different random seeds to ensure statistical reliability and account for Monte Carlo variability in data generation and model fitting. Unless indicated otherwise, each simulation was conducted 3 times, which provided low enough standard errors for coverage rates in most cases ($<1\%$).

3.3 Causal Forest Implementation

This paper analyzes the implementation of Causal Forests using EconML's `CausalForest` class, which makes available the following hyperparameters with the following defaults:

- `n_estimators`: Number of trees in the forest (default: `2500`)
- `criterion`: Function used to measure the quality of a split (default: `'mse'`)
- `max_depth`: Maximum depth of the trees (default: unlimited)

- `min_samples_split`: Minimum number of samples required to split an internal node (default: `10`)
- `min_samples_leaf`: Minimum number of samples required to be at a leaf node (default: `5`)
- `min_weight_fraction_leaf`: Minimum weighted fraction of the total sample weight at a leaf node; if no sample weights were provided, all samples have the same weight (default: `0.0`)
- `min_var_fraction_leaf`: Minimum variance fraction at leaf (default: unset)
- `min_var_leaf_on_val`: Whether the variance constraint is enforced on the validation sample; enabling this breaks honesty (default: `False`)
- `max_features`: Number of features to consider when looking for the best split (default: `'auto'` which is equivalent to 1, or all features)
- `min_impurity_decrease`: Minimum decrease in impurity required to split a node (default: `0.0`)
- `max_samples`: Fraction of the training data used to grow each tree (default: `0.45`)
- `min_balancedness_tol`: Tolerance for how balanced splits must be (default: `0.45`)
- `honestness`: Enables honest splitting by using separate samples for splitting and estimation (default: enabled)
- `inference`: Enables variance estimation for inference on treatment effects (default: enabled)
- `fit_intercept`: Whether to include an intercept in the treatment effect model (default: enabled)
- `subforest_size`: The number of trees within each sub-forest employed in the bootstrap-of-little-bags calculation (default: `4`)

Unless specified otherwise, these default values were used.

3.4 Methodology per sub-question

This segment outlines the methodology corresponding to each sub-question. While beyond the scope of this section, it is important to note that the configurations for the DGP for the second and third sub-questions were chosen based on the results of the first sub-question. Specifically, six pivotal points were selected within the Confounders × Effect Modifiers space: (1, 1), (2, 2), (2, 0), (1, 2), (0, 2), (0, 1). The polynomial degree, number of instruments, and the strength of confounding were set at 3, 1, and 1, respectively.

3.4.1 Sub-question 1 – Data Characteristics Sensitivity

Sobol sampling was used to investigate the parameter space with 1536 (before excluding invalid combinations) samples over the following five parameters:

Parameter	**Range**	**Argumentation for Range**
Polynomial degree	$\{1, 2, 3, 4, 5\}$	These ranges were chosen to reflect low to moderate polynomial degrees
Confounding strength	$(0, 10]$	This spectrum was selected to align with what is typically referred to in the literature as low to high confounding strength
Number of confounders	$\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$	These ranges were chosen to reflect low to medium dimensionality due to computational reasons
Number of instruments	$\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$	
Number of effect modifiers	$\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$	

Configurations are deemed invalid if both the count of instruments and confounders is zero, or if both the count of confounders and effect modifiers is zero.

Subsequently, HDMR with a maximum interaction order of 2 was utilized to evaluate the sensitivity of coverage rates to quantify the contribution of each parameter and their interactions to model performance variability, in order to identify the most influential parameters for further analysis.

To determine practical limits on coverage rate performance, a grid search was performed for the two most significant parameters: expanded range (0–15) for confounders and effect modifiers, while fixing other parameters (polynomial degree=2, confounding strength=1, instruments=1). Subsequently, another grid search was performed on the same parameters but with more resources (5000 trees, 100000 data points) and a sparser grid ($\{0, 3, 9, 12, 15\} \times \{0, 3, 9, 12, 15\}$) to study how coverage rates behave with increased computational resources during training. These grid searches identified a very limited subsection of the search space for which a coverage of more than 60% was achieved, and for which an extensive grid search was performed: 0–4 Confounders × Effect Modifiers while varying polynomial degree $\in \{1, 2, 3\}$, confounding strength $\in \{0.5, 1.0, 1.5, 2.0\}$, and instruments $\in \{0, 1, 2, 3, 4\}$.

3.4.2 Sub-question 2 – Hyperparameter Analysis

For this sub-question, due to the high number of hyperparameters and low computational resources, a full sensitivity analysis with HDMR was not possible. Instead, each hyperparameter was analyzed independently of the others using grid searches for 6 different configurations of the DGP.

Hyperparameters that are known to break the honesty (`min_var_fraction_on_val`, `honestness`), disable confidence intervals (`inference`), or have no effect on confidence intervals (`fit_intercept`), as specified in EconML's documentation [20] were not evaluated. In addition, `min_samples_leaf` and `min_samples_split` were evaluated solely with integer values. This is because they are equivalent to the fractional forms, if the number of samples per leaf or split is divided by the total sample size (in this case, 25000). The hyperparameter `min_weight_fraction_leaf` was not assessed, since without any sample weights, it functions the same as `min_samples_leaf` when given a fraction.

In this sub-question, to guarantee consistent results with minimal standard errors, all experiments were conducted 10 times, unless specified otherwise, with each of the following hyperparameters being examined separately:

1. `criterion` $\in \{$`"mse"`, `"het"`$\}$ with 100 repetitions
2. `max_depth` $\in \{1, 5, 10, 12, 14, 16, 18, 20, 25, 50, 100, 200, 500, 1000,$ `null`$\}$
3. `max_features` $\in \{1, 2, 3, 4\}$ with 100 repetitions
4. `max_samples` $\in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35\}$ and $\{0.4, 0.45, 0.5\}$ with 100 repetitions
5. `min_balancedness_tol` $\in \{0,\ 0.005,\ 0.01,\ 0.015,\ 0.02,\ 0.03,\ 0.04,\ 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3,\ 0.35,\ 0.4,\ 0.45,\ 0.5\}$
6. `min_impurity_decrease` $\in \{0.0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$
7. `min_samples_leaf` $\in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$
8. `min_samples_split` $\in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50\}$
9. `min_var_fraction_leaf` $\in \{$`null`$, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 0.9, 1.0\}$
10. `n_estimators` $\in \{4, 12, 40, 120, 400, 1200, 2400, 4000, 6000, 8000, 10000\}$
11. `subforest_size` $\in \{2, 4, 5, 10, 20, 25, 50, 100\}$

For each hyperparameter, these six configurations were tested:

- Confounders × Effect Modifiers $\in \{(1, 1), (2, 2), (2, 0), (1, 2), (0, 2), (0, 1)\}$
- Number of Instruments = 1
- Polynomial degrees = 3
- Confounding strengths = 1
- Tree count = 2500
- Training sample size = 25000

3.4.3 Sub-question 3 – Tree Count and Sample Size interactions

A grid search was performed for the six points of interest to identify interactions between tree count, sample size, and data characteristics, with all experiments conducted 10 times:

- Confounders × Effect Modifiers $\in \{(1, 1), (2, 2), (2, 0), (1, 2), (0, 2), (0, 1)\}$
- Number of Instruments = 1
- Polynomial degrees = 3
- Confounding strengths = 1
- Tree count $\in \{3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$
- Training sample size $\in \{10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 110000, 120000\}$

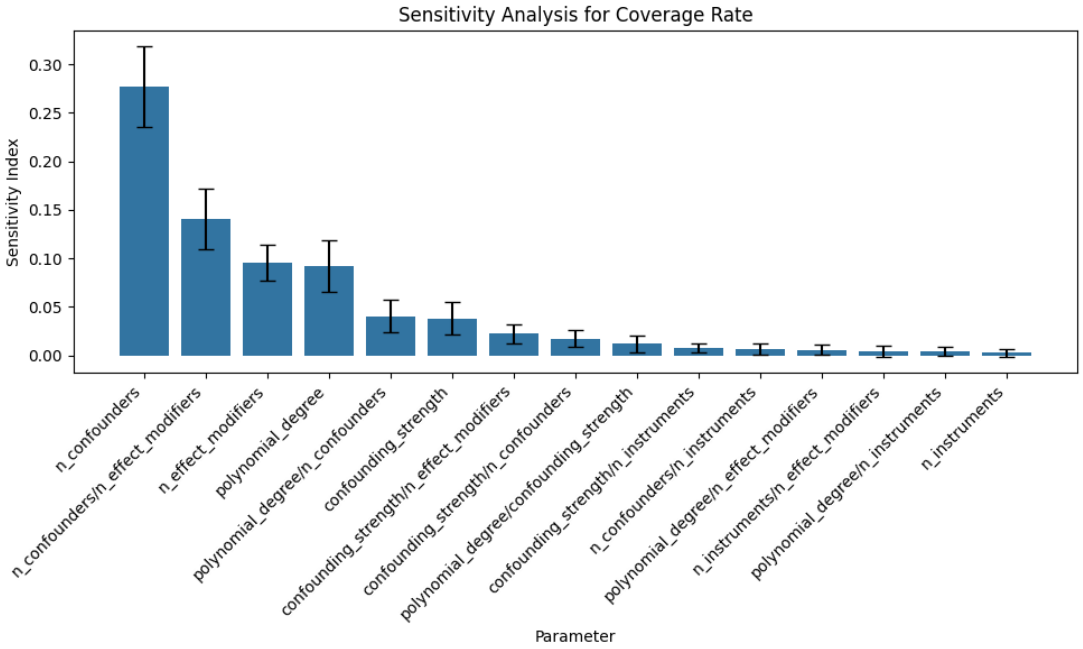

Figure 1: HDMR sensitivity analysis for coverage rate with maximum interaction order of 2. The number of confounders, their interaction with effect modifiers, and effect modifiers are the dominant factors (sensitivity indices $\approx 0.28 \pm 0.04$, $\approx 0.14 \pm 0.03$, and $\approx 0.10 \pm 0.02$), followed by polynomial degree, its interaction with the number of confounders, and confounding strength ($\approx 0.09 \pm 0.03$, $\approx 0.04 \pm 0.02$, and $\approx 0.04 \pm 0.01$), while all other parameters and interactions contribute minimally to coverage rate variability (< 0.02).

The starting thresholds of 3000 trees and 10000 datapoints were selected to avoid the unstable region for lower values due to considerable Monte Carlo noise.

4 Results

4.1 Sub-question 1 – Data Characteristics Sensitivity

Following the initial Sobol sampling, the HDMR analysis identified the number of confounders and effect modifiers as the most influential factors with respect to the coverage rate of the confidence intervals, as shown in Figure 1.

As specified in the methodology section, based on the HDMR results, a series of grid searches was performed on confounders × effect modifiers, the two most influential parameters.

- **Grid Search 1**: The first grid search that was performed on an extended area (0–15 confounders × 0–15 effect modifiers) indicated that the coverage rates decrease considerably, falling below 60% when the combined number of confounders and effect modifiers exceeds 4, as shown in Table 1. A complete heat map can be seen in Figure A.3.

- **Grid Search 2**: This grid search revealed that doubling the number of trees and increasing the training data fourfold led to improvements under 7% when compared to the previous grid search's results, as seen in Figure A.4.

- **Grid Search 3**: The third grid search revealed that even for simple configurations (confounding strength = 0.5 and polynomial degree = 1), the coverage rate drops below 80% as soon as the number of confounders and effect modifiers exceeds 4, as shown in Table 1.

	Degree 2, Conf. Strength 1		Degree 1, Conf. Strength 0.5	
Confounders/Modifiers	**Avg. Coverage**	**Max Coverage**	**Avg. Coverage**	**Max Coverage**
1	0.94	0.94	0.95	0.95
2	0.92	0.93	0.94	0.94
3	0.76	0.82	0.91	0.92
4	0.64	0.66	0.86	0.91
5	0.51	0.54	0.77	0.78
6+	< 0.50	< 0.50	< 0.70	< 0.70

Table 1: Coverage Rates by Polynomial Degree and Confounding Strength

4.2 Sub-question 2 – Hyperparameter Analysis

The hyperparameter analysis revealed varying behaviors across parameters. Table 2 summarizes the key statistics for the impact of each hyperparameter on coverage rates. Detailed plots for all experiments can be found in figs. A.5 to A.15. The following paragraphs summarize notable behaviors of the average coverage rates of the 6 configurations.

Parameter	**Min**	**Max**	**Mean**	**Range**	**Std. Err.**
criterion	0.7328	0.7428	0.7378	0.0100	0.0047
max_depth	0.0860	0.7525	0.6296	0.6665	0.0018
max_features	0.5621	0.7230	0.6295	0.1609	0.0014
max_samples	0.5883	0.7594	0.6838	0.1711	0.0019
min_balancedness_tol	0.0451	0.7561	0.5783	0.7111	0.0016
min_impurity_decrease	0.2211	0.7328	0.4175	0.5117	0.0027
min_samples_leaf	0.5560	0.7530	0.6762	0.1971	0.0014
min_samples_split	0.6264	0.7530	0.7153	0.1267	0.0013
min_var_fraction_leaf	0.1517	0.7413	0.6216	0.5896	0.0041
n_estimators	0.6786	0.8508	0.7602	0.1723	0.0019
subforest_size	0.7285	0.7480	0.7388	0.0195	0.0019

Table 2: Statistics of Coverage Rate Across Different Parameters

The coverage rates for `max_depth` began at a low point, then steadily rose until the average stabilized at approximately 0.74 for max_depth = 20.

For each rise in the `max_features` value, the coverage rates have either stayed constant or shown improvement.

The `max_samples` parameter demonstrated a positive correlation with coverage rates. Higher values consistently improved performance, particularly when exceeding 0.4. As the value increased from 0.4 to 0.5, the coverage rate and the width of the confidence interval increased by 0.0375 and 0.2522, respectively. A setting of 0.45 achieved the 0.7422 coverage rate and the width of 2.6290, while 0.5 delivered the maximum coverage rate of 0.7594 with broader intervals of 2.7527.

The `min_balancedness_tol` parameter showed that increasing values improved coverage rates. Coverage initially increased for values from 0 to 0.1, but then stabilized at approximately 0.74. Between 0.1 and 0.5, the coverage rates remained relatively steady while the width of the confidence interval decreased from 3.1726 to 2.5676.

The `min_impurity_decrease` parameter exhibited a negative relationship with coverage rates. As the values increased from 0 to 1, the coverage rates decreased monotonically, ranging from 0.732 to 0.221.

For `min_samples_leaf`, optimal performance was observed with values up to 5, beyond which there was a marked decline in coverage rates. Performance increased marginally from 1 to 5 (difference: 0.012), followed by a noteworthy drop from 5 to 50 (difference: 0.197). Meanwhile, the mean squared error almost doubled within the same interval (5 to 50). At an optimal value of 5, the coverage rate was 0.7530.

For `min_samples_split`, the coverage rate remained quite steady, fluctuating between 0.73 and 0.75 up to 14 samples, after which it experienced a continuous drop to 0.64 for 50 samples.

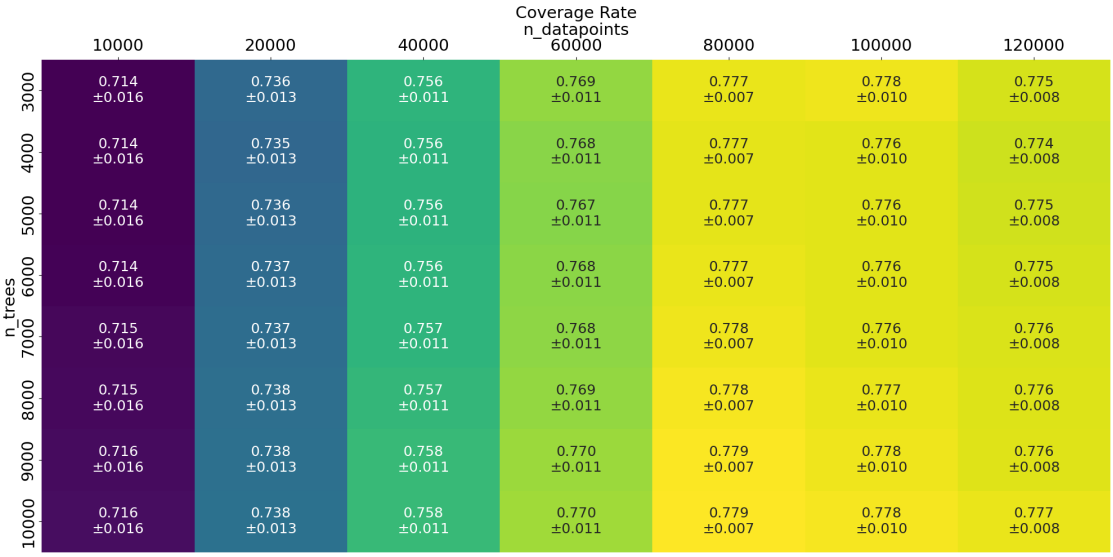

Figure 2: Interaction effects of tree count and training sample size on model performance averaged across six combinations of confounders and effect modifiers.

The `min_var_fraction_leaf` parameter showed stable behavior until a configuration-specific threshold was reached, beyond which performance sharply declined. The variability before reaching these thresholds was minimal.

The `n_estimators` parameter showed complex dynamics across the different metrics. MSE stabilized relatively quickly between 120 and 400 trees. CI width followed a U-shaped pattern: starting at 4.4 for 4 trees, increasing to 4.88 for 12 trees, then decreasing monotonically until stabilizing around 2.63 at approximately 2400 trees. The coverage rates exhibited similar behavior, starting at 0.6786 for 4 trees, peaking at 0.8505 for 40 trees, then relatively stabilizing around 0.7331 for 2400 trees, with diminishing returns reaching 0.7335 at 10000 trees.

4.3 Sub-question 3 – Tree Count and Sample Size interactions

Since all six configurations exhibited similar behavior, results were averaged across configurations and displayed as a heatmap in Figure 2. On average, increasing the training size from 10000 to 120000, a twelve-fold increase, resulted in an improvement of $\approx 6.1\%$. However, increasing the number of trees from 3000 to 10000, a 3.3-fold increase resulted in marginal improvements of less than 0.2%.

5 Discussion

5.1 Sub-question 1 – Data Characteristics Sensitivity

The sensitivity analysis using HDMR indicated that confounders and effect modifiers have the most notable impact on the coverage of confidence intervals, followed by polynomial degree and confounding strength. Unsurprisingly, the number of instruments minimally affects the coverage rate, likely because partitioning by instruments does not offer any enhancements to the splitting criterion and, consequently, does not get selected. Although these findings are limited to the ranges studied, they offer valuable practical advice to practitioners by clearly prioritizing the importance of various factors when under comparable constraints.

Conversely, a concept that can be broadly applied to all datasets with polynomial treatment effects is that coverage rates decline considerably once the combined total of confounders and effect modifiers surpasses 4, thus identifying a practical threshold not yet reported in academic literature. This threshold seems robust, as enhancing computing power (by doubling the number

of trees and quadrupling the dataset size to 5000 and 100000, respectively) resulted in improvements under 7%. Combined with the findings of Section 4.3, which show that both the number of trees and the data points face notable diminishing returns, this implies an intrinsic limitation rather than a computational one. These results offer practical advice: researchers with more than four combined confounders and effect modifiers should anticipate considerably reduced reliability of the confidence intervals and consider alternative methodologies.

Furthermore, as mentioned in the methodology, these findings led to the decision to concentrate the following analyzes on 6 specific confounders $\times$ effect modifiers $\in \{(1, 1), (2, 2), (2, 0), (1, 2), (0, 2), (0, 1)\}$. This choice was driven by the rapid degradation observed, thus opting to target regions nearer to the nominal coverage rates for more useful insights. Furthermore, confounding strength, polynomial order, and instrument count were fixed at 1, 3, and 1, respectively, given computational constraints and their relatively small impact on variance.

5.2 Sub-question 2 – Hyperparameter Analysis

Table 3 offers a comprehensive analysis of each parameter, giving practical recommendations based on their observed effects. These parameters are systematically classified by their impact range-wise: very low (0.01-0.05), low (0.12-0.20), medium (0.50-0.52), high (0.58-0.60), and very high (0.66-0.72).

Table 3: Hyperparameter discussion regarding Confidence Intervals

Parameter	Impact	Observed Effect	Interpretation and Recommendation
`criterion`	Very Low	Minor variation; effects likely due to Monte Carlo noise.	No meaningful impact observed on coverage. Default (mse) is sufficient; no tuning needed.
`max_depth`	Very High	Coverage increases with tree depth, plateauing around depth 20. Shallow trees perform poorly.	Shallow trees underfit, reducing confidence interval reliability. For best results in general, leave unset.
`max_features`	Low	Higher values improve or maintain coverage; no observed degradation.	Including more features when splitting improves coverage by allowing better modeling of effect heterogeneity. The default value, which includes all available features, should be generally safe.
`max_samples`	Low	Higher values ($\geq$ 0.4) consistently improve coverage; peak at 0.5.	Larger sample sizes per tree reduce variance in estimation, leading to more stable confidence intervals. Recommended to increase to 0.5 for best coverage rate performance.
`min_balancedness_tol`	Very High	Strong positive effect; coverage rates plateau after 0.1, although widths continue to decrease.	Improves coverage rates by ensuring balanced splits. Setting this value near the upper limit seems best for confidence intervals overall. Set to 0.5 for optimal coverage and width performance.
`min_impurity_decrease`	Medium	Higher values degrade coverage. Even small increases from 0 hurt performance.	This parameter prevents beneficial splits, forcing trees to stop early and creating large leaves. This violates Specification 1 from Athey *et al.* [4] by allowing trees to refuse splitting on any feature when impurity gains are below the threshold, effectively making the split probability 0 rather than bounded below by some $\pi > 0$ as required, invalidating the theoretical foundation for confidence intervals. Keep at default (0.0) to avoid degrading inference quality.
`min_samples_leaf`	Low	Best coverage at small values (≤ 5); higher values sharply reduce performance.	Larger leaves average over heterogeneous subgroups, diluting treatment effects, resulting in worse performance. Default (5) is adequate; no tuning necessary.
`min_samples_split`	Low	Minimal effect until 15 samples; then moderate decline in performance.	Avoid large values as they reduce tree depth and heterogeneity modeling. Default (10) is adequate; no tuning necessary.

Continued on next page

Table 3 – continued from previous page

Parameter	Impact	Observed Effect	Interpretation and Recommendation
`min_var_fraction_leaf`	High	Threshold behavior shows a sharp decline after configuration-specific limits with minimal variability beforehand.	This likely stems from mechanisms similar to those in `min_impurity_decrease`, with the initial low variability due to less constraining power for low values. Leave unset for best performance.
`n_estimators`	Low	Coverage rate unstable until approximately 2400 trees when the theoretical "sufficiently large" requirement is practically achieved; Diminishing returns beyond this point	According to Wager and Athey [3] and Athey *et al.* [4], it is necessary for the number of trees to be high enough so that the Monte Carlo noise becomes insignificant for valid confidence intervals. Use $\geq$ 2400 trees for best performance.
`subforest_size`	Very Low	Minor variation; effects likely due to Monte Carlo noise.	There was no noteworthy impact detected on coverage. Default (4) is sufficient; no tuning needed.

5.3 Sub-question 3 – Tree Count and Sample Size interactions

No substantial interactions between the quantity of trees and the size of the training dataset have been detected. Additionally, it appears that upon reaching the sufficiently large number of trees, required in order for the Monte Carlo noise to become negligible, further increases lead to only very minor improvements of less than 0.2%. A similar pattern is observed with respect to the number of data points; specifically, within the 80000 to 120000 data points interval, the coverage rates stabilized around 77.6% with variations below 0.2%.

The evidence indicates that although expanding the number of trees and the sample size improves the coverage rate, these advantages wane over time. Since there are no interaction effects, the number of trees and sample size can be optimized individually. This independence, coupled with the diminishing returns, enables practitioners to fine-tune each parameter separately according to their computational limits and desired coverage rate performance, without needing to explore exhaustive parameter combinations.

6 Responsible Research

6.1 Ethical Implications

This study tackles a significant deficiency in understanding the dependability of causal inference techniques, which directly affects evidence-based decisions in healthcare, policy, and other crucial areas. By pinpointing the circumstances under which causal forest confidence intervals lose their reliability, this research fulfills an ethical obligation to prevent overconfident causal assertions that may result in detrimental outcomes. However, it is important to note that this study was conducted on synthetic datasets, which may not align with real-world datasets. Thus, in order to safely use these findings with real observational data, it is imperative to first validate this study with real-world data.

6.2 Reproducibility

Reproducibility-wise, all experiments can be fully replicated due to the fact that all of the experimental setups (hyperparameters, data characteristics) and methodology are documented in this paper. In addition, the codebase is published online [21], uses readily available libraries (for example, EconML and SaLib), and contains the full configurations used for all experiments, reducing the act of performing the experiments to just running the right command. In addition, to ensure statistical reliability, all experiments were run multiple times (at least 3 times, but mostly 10 times, in some cases even 100 times), with different random seeds. The inclusion of deterministic quasirandom sampling (Sobol sequences) improves reproducibility relative to purely random sampling methods, allowing for the exact reproduction of parameter space exploration

across different test runs. Finally, all significant results include standard errors or confidence intervals.

7 Limitations and Future Work

This research provides substantial takeaways on the reliability of the confidence intervals in GRF causal forests, but several limitations affect the generalizability of the findings. The research relies exclusively on synthetic data from a specific data-generating process that provides a polynomial treatment effect, which may not capture the full complexity of real-world observational data. The investigation was constrained to specific parameter ranges (confounders, effect modifiers, and instruments up to 8, polynomial degrees up to 5, and confounding strength up to 10) due to computational limitations. Higher-dimensional scenarios common in modern applications have not been rigorously analyzed, although the limit of 4 combined confounders and effect modifiers for reliable coverage would probably be even more restrictive in complex contexts. The analysis was limited to binary treatments, while a considerable amount of applications deal with continuous or multi-valued treatments. Furthermore, because of computational limitations, hyperparameters were investigated individually rather than holistically, which could result in overlooking significant interaction effects.

It is recommended for future research on Causal Forest's confidence intervals to attempt to rectify these shortcomings in the following ways. Firstly, validation should incorporate synthetic datasets with a wider range of treatment effect functions, such as exponential, logarithmic, and discontinuous effects, rather than limiting to polynomials, to validate the generalizability of these findings. Secondly, investigations should expand beyond binary treatments to include continuous and multi-valued interventions. Thirdly, a thorough analysis of hyperparameter interactions should be conducted to enhance the independent analysis done here, capturing key interaction effects that influence coverage reliability. In addition, developing adaptive hyperparameter tuning algorithms specifically for coverage rate improvement will equip practitioners with data-driven methods to achieve optimal confidence interval coverage specific to their applications. Finally, and most importantly, these findings should be validated on real observational data.

8 Conclusions

This research addressed a critical deficiency in understanding confidence interval reliability in causal forests by systematically investigating how data characteristics and hyperparameters affect coverage rates. The main research question focuses on determining the factors (data characteristics and hyperparameters) that affect the actual coverage rates of confidence intervals and the optimal hyperparameters to optimize them.

The most substantial finding is the identification of a practical threshold for reliable confidence interval coverage: when the combined number of confounders and effect modifiers exceeds 4, coverage rates decline dramatically below 80% even for the simplest treatment effect function. This threshold appears robust, as increasing computational resources provided only marginal improvements. This limitation was not previously documented in the literature and has important ramifications for practitioners working with complex observational data with a polynomial treatment effect.

The hyperparameter analysis also uncovered some notable findings. The most impactful parameters were maximum depth and minimum balancedness tolerance, which caused the coverage rates throughout their range to vary by 66% and 71%, respectively. Key recommendations include using at least 2400 trees to satisfy theoretical requirements for valid confidence intervals, setting maximum samples per tree to 0.5 and minimum balancedness tolerance to 0.5 for optimal coverage, leaving maximum depth unset to avoid underfitting, and maintaining minimum impurity decrease at 0.0 to prevent degradation of inference quality. Several parameters showed minimal impact (splitting criterion, subforest size) and can be left at default values without affecting coverage rates.

The interaction analysis between tree count and sample size revealed diminishing returns as the parameters increase, thus providing clear cost-benefit guidance: practitioners can identify optimal

resource allocation points where further increases in trees or training data yield improvements that are too small to justify the additional computational cost. The absence of interaction effects between these parameters simplifies optimization by allowing independent tuning of tree count and sample size rather than requiring joint optimization.

The identification of coverage rate degradation beyond 4 combined confounders and effect modifiers, along with the hyperparameter recommendations, may help practitioners arrive at more informed conclusions when applying causal forests to their observational data.

With the rapid advancement and increasing application of causal inference in healthcare, policy, and business decision-making, it is essential for practitioners to discern when confidence intervals are reliable and how to optimize them. This study lays critical groundwork by offering empirically based insights that connect theoretical assurances with practical dependability while establishing important practical guidelines that, although requiring validation with real-world datasets for broader applicability, can help prevent overconfident causal claims and ultimately serve the goal of more reliable and responsible decision-making.

A Additional Charts

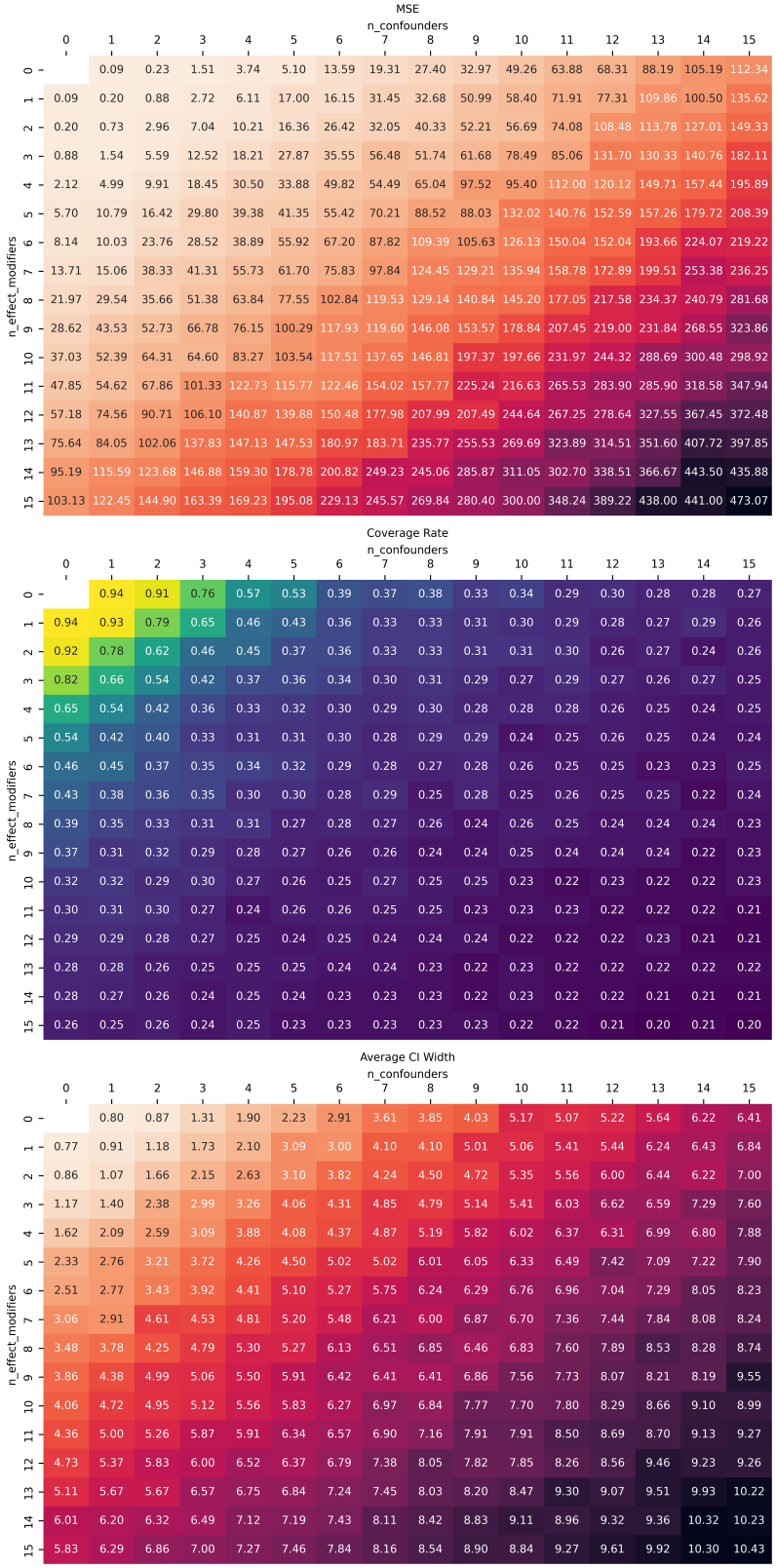

Figure A.3: Performance metrics across varying numbers of confounders (x-axis) and effect modifiers (y-axis) using 2500 trees and 25000 data points. Coverage rates deteriorate more rapidly than MSE as the number of confounders and effect modifiers increases.

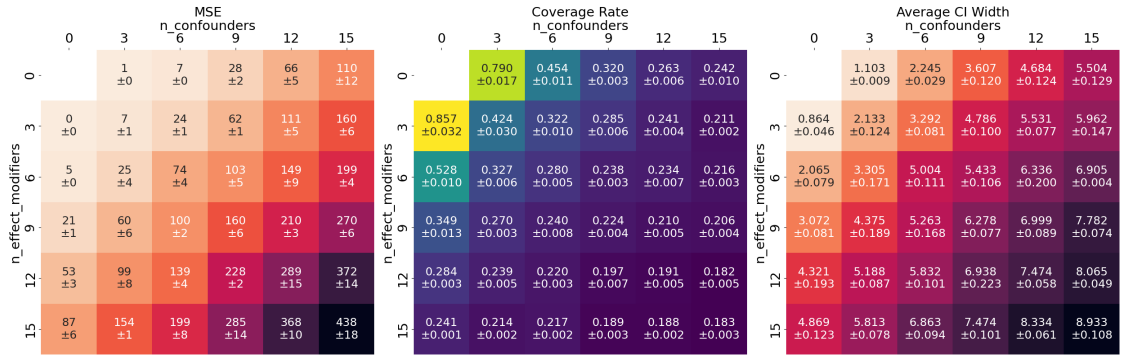

Figure A.4: Performance metrics from a grid search using 5000 trees and 100000 data points across the same parameter space. Despite the fourfold increase in training data and doubled number of trees, improvements over Figure A.3 were minimal (under 7%). This plot reports values ± std. deviation.

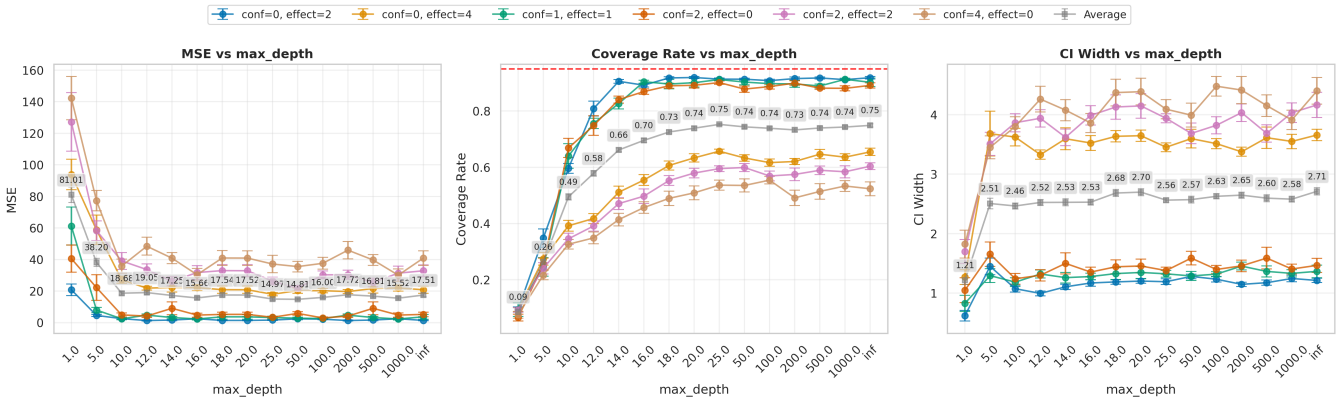

Figure A.5: `max_depth` analysis

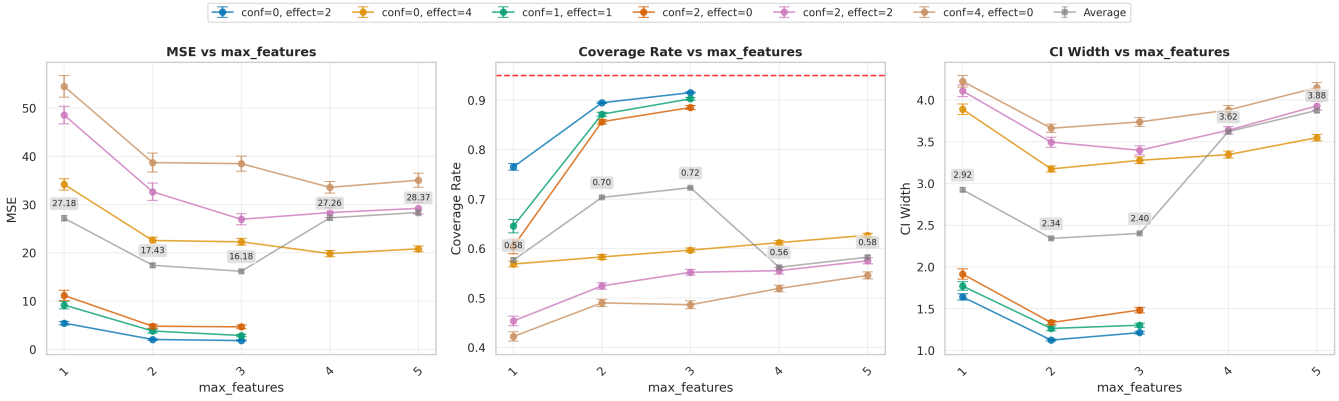

Figure A.6: `max_features` analysis

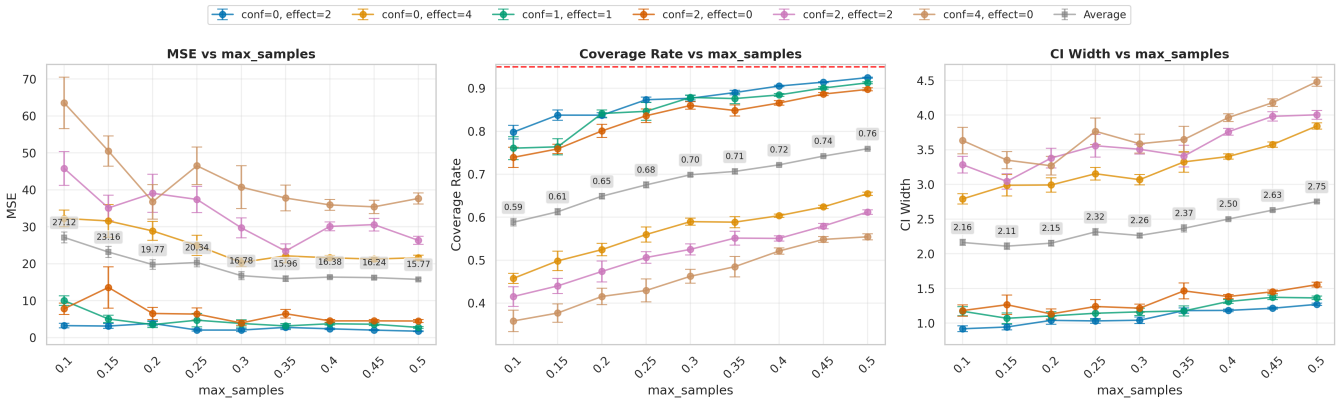

Figure A.7: `max_samples` analysis

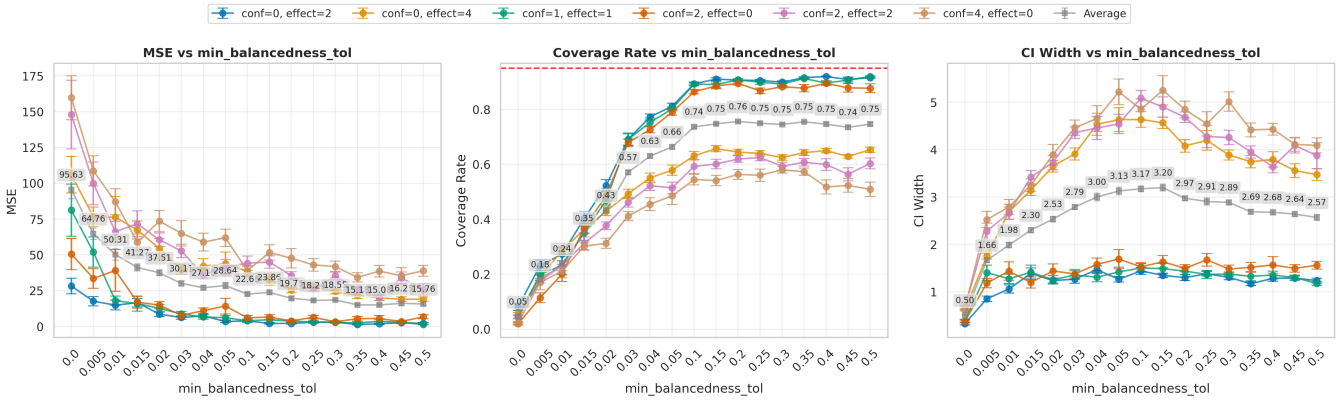

Figure A.8: `min_balancedness_tol` analysis

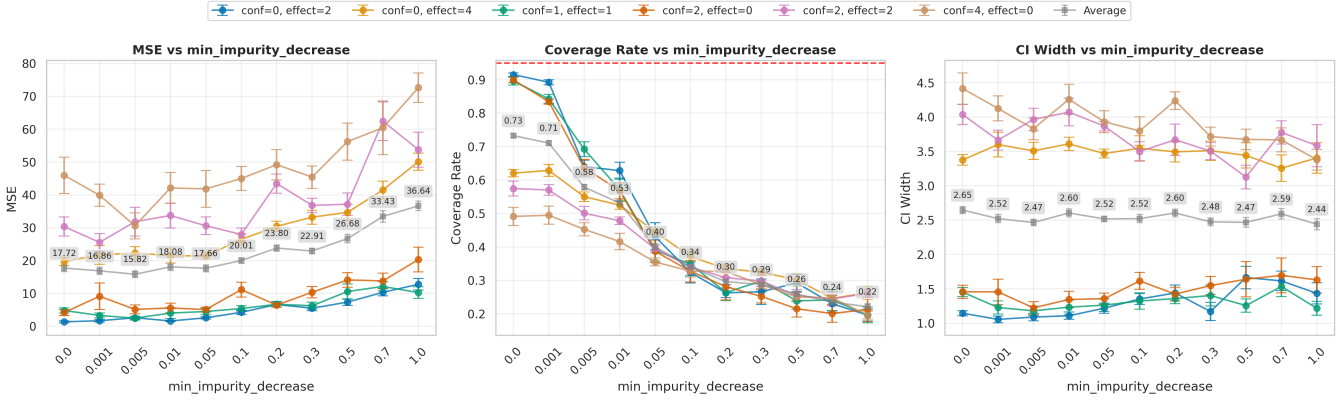

Figure A.9: `min_impurity_decrease` analysis

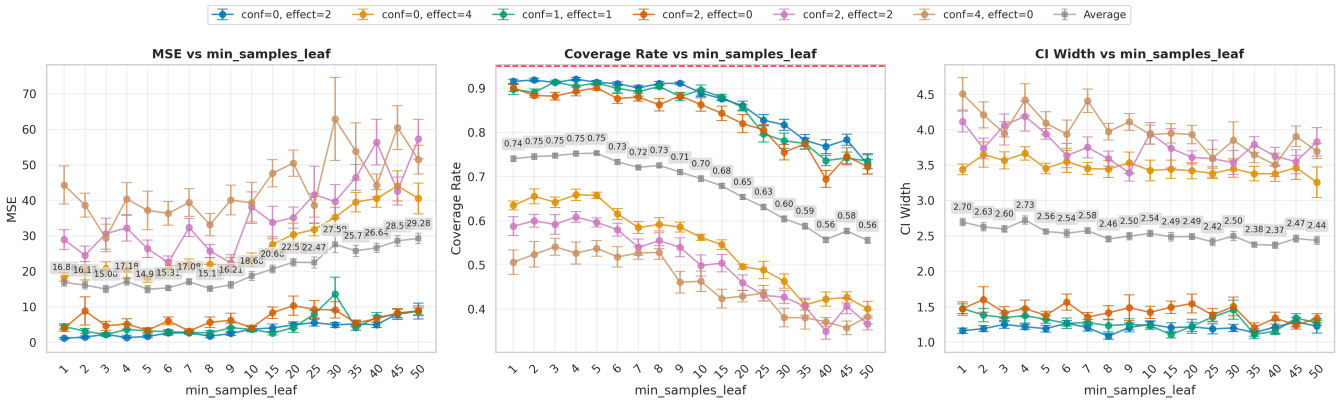

Figure A.10: `min_samples_leaf` analysis

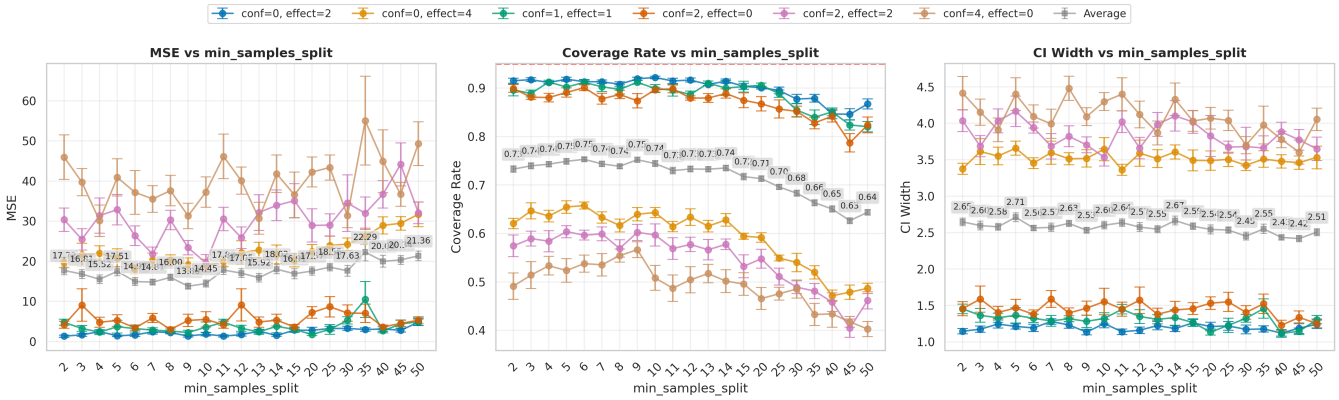

Figure A.11: `min_samples_split` analysis

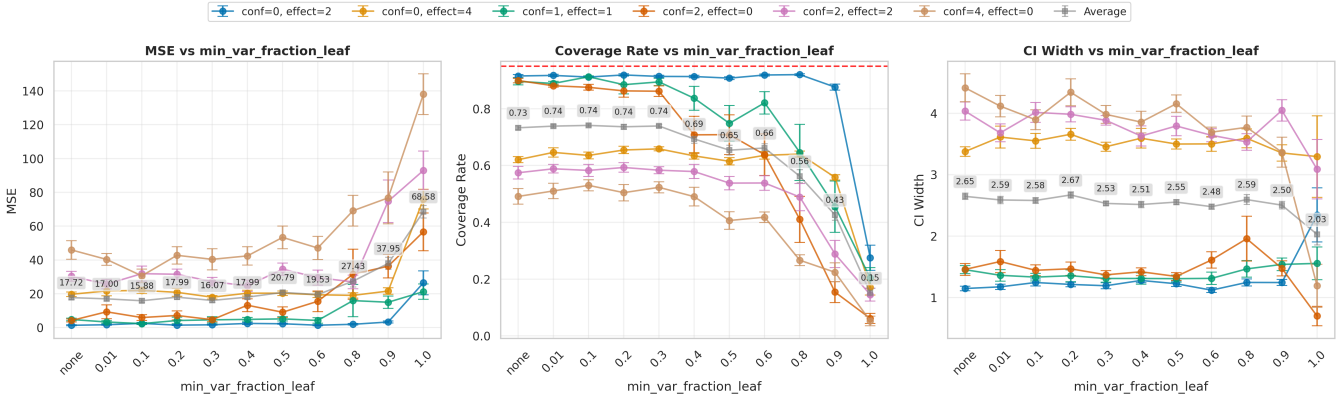

Figure A.12: `min_var_fraction_leaf` analysis

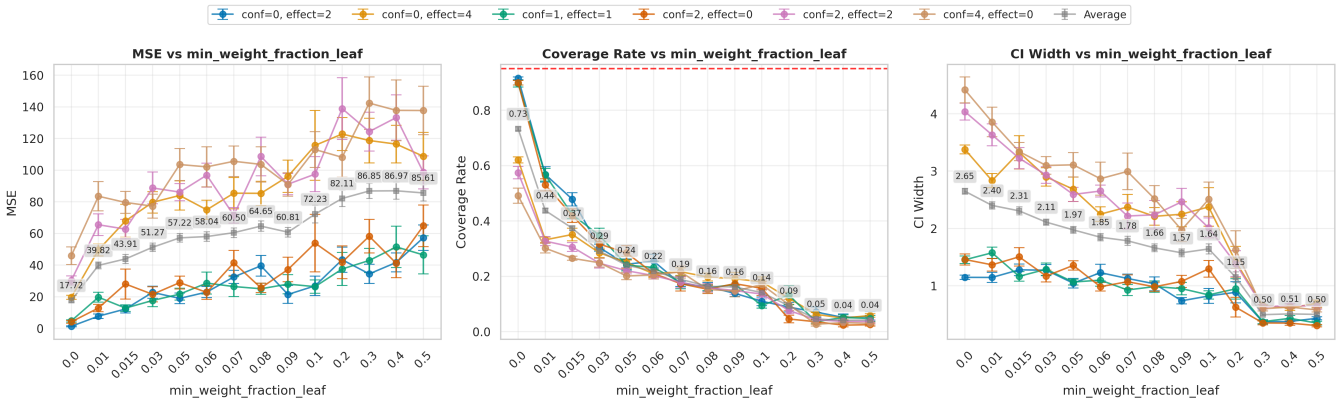

Figure A.13: `min_weight_fraction_leaf` analysis

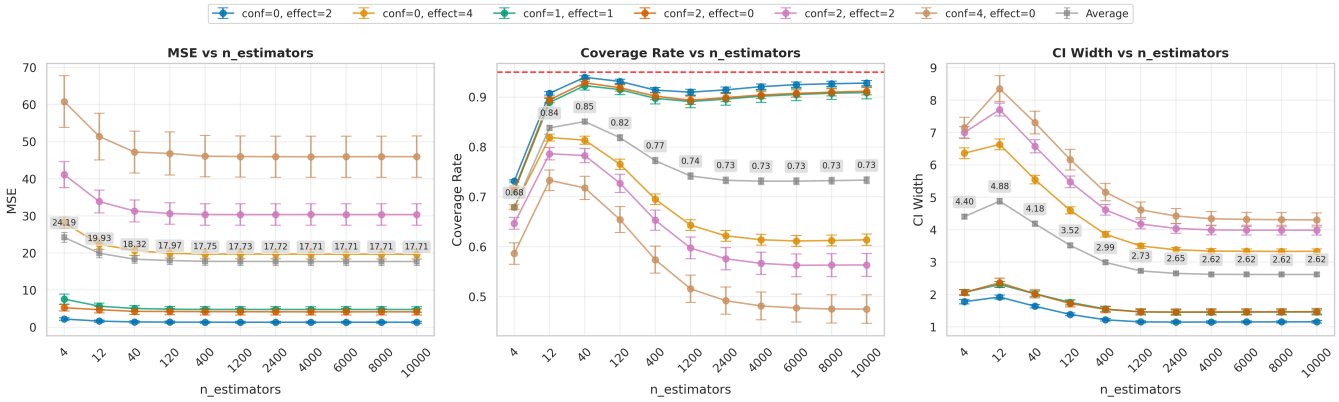

Figure A.14: `n_estimators` analysis

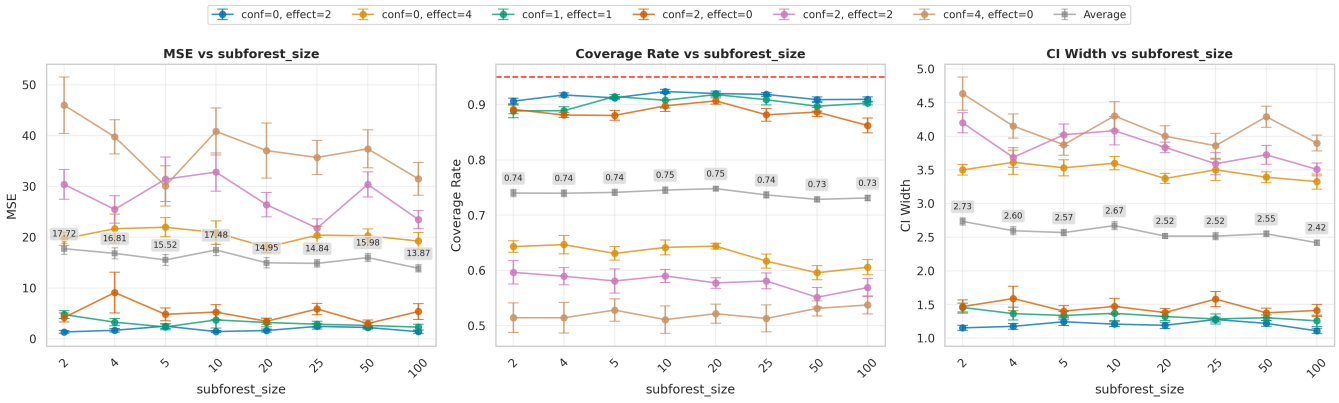

Figure A.15: `subforest_size` analysis

References

[1] S. Feuerriegel *et al.*, "Causal machine learning for predicting treatment outcomes", *Nature Medicine*, vol. 30, no. 4, pp. 958–968, 2024.

[2] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests", *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[4] S. Athey *et al.*, "Generalized random forests", *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019. DOI: 10.1214/18-AOS1709. [Online]. Available: https://doi.org/10.1214/18-AOS1709.

[5] Y. Saito and S. Yasui, "Counterfactual cross-validation: Stable model selection procedure for causal inference models", in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, Vienna, Austria: PMLR, 2020.

[6] L. Lei and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 83, no. 5, pp. 911–938, 2021.

[7] S. Wager *et al.*, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife", *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1625–1651, 2014.

[8] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies", *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[9] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press, 2015.

[10] J. Sexton and P. Laake, "Standard errors for bagged and random forest estimators", *Computational Statistics & Data Analysis*, vol. 53, pp. 801–811, 2009. DOI: 10.1016/j.csda.2008.08.007.

[11] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals", *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 4, pp. 86–112, 1967.

[12] J. Dick *et al.*, "High-dimensional integration: The quasi-monte carlo way", *Acta Numerica*, vol. 22, pp. 133–288, 2013.

[13] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices", *Computer Physics Communications*, vol. 145, no. 2, pp. 280–297, 2002. DOI: 10.1016/S0010-4655(02)00280-1.

[14] S. Tarantola *et al.*, "A comparison of two sampling methods for global sensitivity analysis", *Computer Physics Communications*, vol. 183, no. 5, pp. 1061–1072, 2012.

[15] B. Sudret *et al.*, "Quasi random numbers in stochastic finite element analysis–application to global sensitivity analysis", in *International Conference on Applications of Statistics and Probability in Civil Engineering*, 2007.

[16] M. Renardy *et al.*, "To sobol or not to sobol? the effects of sampling schemes in systems biology applications", *Mathematical Biosciences*, vol. 337, p. 108 593, 2021. DOI: 10.1016/j.mbs.2021.108593. [Online]. Available: https://doi.org/10.1016/j.mbs.2021.108593.

[17] H. Rabitz and Ö. F. Aliş, "General foundations of high-dimensional model representations", *Journal of Mathematical Chemistry*, vol. 25, no. 2-3, pp. 197–233, 1999. DOI: 10.1023/A:1019188517934.

[18] G. Li *et al.*, "Global sensitivity analysis for systems with independent and/or correlated inputs", *The Journal of Physical Chemistry A*, vol. 114, no. 19, pp. 6022–6032, 2010. DOI: 10.1021/jp9096919.

[19] J. Herman and W. Usher, *SALib HDMR API Reference*, https://salib.readthedocs.io/en/latest/api.html\#high-dimensional-model-representation/, [Online; accessed 06-June-2025]. [Online]. Available: https://salib.readthedocs.io/en/latest/api.html%5C#high-dimensional-model-representation (visited on 06/06/2025).

[20] Microsoft EconML Team, *Econml causalforest documentation*, https://econml.azurewebsites.net/_autosummary/econml.grf.CausalForest.html, [Online; accessed 13-June-2025], 2024. [Online]. Available: https://econml.azurewebsites.net/_autosummary/econml.grf.CausalForest.html (visited on 06/13/2025).

[21] R. Iordan, *Causal forest research*, GitHub repository, 2025. [Online]. Available: https://github.com/09Ria09/causal-forest-research.