Sequential Bootstrap for Out-of-Bag Error Estimation: A Simulation-Based Replication and Stability-Oriented Refinement

Cheng Peng

Abstract

Bootstrap resampling is the foundation of many ensemble learning methods, and Out-of-Bag (OOB) error estimation is by far the most popular way to get an internal assessment of generalization performance. In the standard multinomial bootstrap, the number of distinct observations in each resample is random. Although this source of variability exists, it has seldom been separately studied to see how much it really affects OOB-based quantities. To fill this gap, we investigate Sequential Bootstrap (a resampling method that forces every bootstrap replicate to contain exactly the same number of distinct observations) and treat it as a controlled modification of the classical bootstrap within the OOB framework. We carefully reproduce Breiman's five original OOB experiments on both synthetic and real-world datasets and repeat everything under many different random seeds. Our results show that switching from the classical bootstrap to Sequential Bootstrap almost does not change accuracy-related metrics, but it does bring measurable (and data-dependent) reductions in several variance-related measures. Therefore, Sequential Bootstrap should not be regarded as a new method to improve predictive performance, but rather as a tool that helps us understand how the randomness in the number of distinct samples contributes to the overall variance of OOB estimators. This work offers a reproducible setting for studying the statistical properties of resampling-based ensemble estimators and provides useful empirical evidence for future theoretical work on variance decomposition in bootstrap-based systems.

Keywords: sequential bootstrap, out-of-bag error, bagging, tree ensembles, bootstrap resampling

1. Introduction

Ensemble methods based on resampling, especially bootstrap aggregation (bagging), have become very important in modern predictive modeling and in the study of algorithmic stability. One of the most attractive features of bagging is the Out-of-Bag (OOB) estimator. This estimator gives us an internal and data-efficient way to estimate generalization error: for each observation, we only average the predictions from those base learners that did not see this observation during training. Since Breiman first proposed it, OOB estimation has been widely applied in practice. For tree-based models, people often treat it as a reliable and nearly unbiased substitute for a separate test set or cross-validation.

The classical bootstrap that is used in OOB draws n samples with replacement from the original dataset of size n. Although the average number of distinct observations in each resample is well known (approximately $0.632n$ (Efron and Tibshirani 1994)),the exact number is still random. This randomness in the number of distinct observations will affect the training sets, then affect the tree structures, split points and finally the predictions. However, even though there has been a lot of theoretical and empirical research on bootstrap-based ensemble methods, very few studies have tried to separately investigate how this particular source of variability influences OOB estimation.

Sequential Bootstrap (SB) offers a way to make the number of distinct observations in each bootstrap replicate exactly the same. Instead of drawing a fixed number of samples (which is what the classical bootstrap does), it keeps sampling with replacement until a pre-specified number of unique observations is reached (Rao, Pathak, and Koltchinskii 1997). This change greatly reduces the variance of the distinct-sample count across different replicates, while the probability that any single observation is included stays almost the same as in the classical case. Although people have studied Sequential Bootstrap before as a general resampling technique, nobody has looked at what happens when we use it together with Out-of-Bag estimation. In particular, we still do not know whether forcing the number of distinct observations to be constant will affect accuracy-related measures, variance-related measures, or the overall stability of OOB error across different datasets and experimental settings.

Because of this gap in the literature, we carry out a carefully controlled experiment: we keep everything in the learning pipeline exactly the same (the same tree model, the same loss function, the same hyperparameters), and the only thing we change is replacing the classical multinomial bootstrap with Sequential Bootstrap. We fully reproduce Breiman's five original OOB experiments (Breiman 1996b) using both synthetic datasets and real datasets, and we repeat all experiments many times with different random seeds. Through these experiments, we want to see how stabilizing the distinct-sample count influences both accuracy-oriented indicators and variance-oriented indicators. Importantly, we do not treat Sequential Bootstrap as a new method that is supposed to

give better prediction accuracy. Instead, we treat it as a deliberate perturbation tool that helps us understand how the randomness coming from varying distinct-sample counts affects the statistical properties of the OOB estimator.

This study contributes to the literature in the following ways:

1. **Variance-structure perspective**: We provide a targeted examination of distinct-sample-count variability as a structural component of OOB-based estimators.
2. **Controlled perturbation framework**: We introduce a reproducible evaluation design in which SB serves as a stability-oriented resampling operator while preserving model, loss, and tuning specifications.
3. **Empirical characterization**: We show that accuracy-oriented metrics remain effectively invariant, while several variance-sensitive diagnostics exhibit measurable but data-dependent changes, illustrating how OOB interacts with resampling stability.

Our results show clearly how the choice of resampling scheme can affect the internal variability of OOB estimation. At the same time, they give useful empirical evidence that can serve as a solid starting point for future theoretical studies on variance decomposition in bootstrap-based ensemble methods.

2. Related Work & Motivation

2.1 Out-of-Bag estimation (OOB)

Bootstrap aggregation (bagging) builds an ensemble of base learners, each trained on a bootstrap sample drawn from the original dataset (Breiman 1996a). Out-of-Bag (OOB) estimation is a very popular technique that allows us to evaluate the predictive performance without needing an extra held-out validation set (Breiman 1996b). The basic idea is quite simple: for every observation in the training data, we only use the predictions from those trees that did not include this observation in their bootstrap sample, and then we average these predictions to get an internal estimate of the generalization error. This OOB error has been shown to be a good approximation of true test error in many cases, especially when the base learners are decision trees (Breiman 2001). Many previous studies have looked at how reliable OOB estimation is in practice, how much bias it has, and how fast it can be computed (Breiman 1996b). Researchers have also compared it with cross-validation on real datasets and usually found that OOB performs similarly well or even better in many situations (Janitza and Hornung 2018).

However, almost all of these existing studies treat the classical multinomial bootstrap as something fixed that does not need to be questioned. Of course, people have proposed many other

resampling methods for ensemble learning, for example subsampling (Politis and Romano 1994), different weighted bootstrap schemes, or stratified and class-aware sampling strategies (Barbe and Bertail 2012), and stratified or class-sensitive sampling (Bickel and Freedman 1984). But these methods are usually designed to change the bias-variance tradeoff, to handle imbalanced classes, or to serve some other modeling purpose. They do not specifically try to separate and study the variability that comes only from the random number of distinct observations in each bootstrap replicate. To the best of our knowledge, no previous work has carefully investigated how this particular source of variability affects the behaviour of OOB-based performance evaluation.

2.2 Sequential Bootstrap

Sequential Bootstrap (SB) is a resampling method that keeps drawing samples with replacement until a fixed number of distinct observations have been collected, instead of stopping after exactly n draws like the classical bootstrap does (Rao, Pathak, and Koltchinskii 1997). Because of this design, the number of unique observations becomes almost the same in every replicate, which greatly reduces its variance and can be understood as making the effective sample size more stable for each base learner (Babu, Pathak, and Rao 2000). Earlier studies on Sequential Bootstrap mainly looked at its theoretical sampling properties, the marginal probability of including each observation, and how it might be used in simulation studies or statistical inference (Babu, Pathak, and Rao 1999).

When people compare different resampling methods in ensemble learning, they usually care about getting higher prediction accuracy, changing the bias-variance tradeoff, or doing something different with the features (for example random feature selection in Random Forests) (Ho 1998). Almost none of them pay attention to whether making the number of distinct observations stable actually matters. As far as we know, there is no published work - neither empirical nor theoretical - that has ever used Sequential Bootstrap together with bootstrap-based ensemble predictors or with Out-of-Bag estimation.

2.3 Motivation for Combining SB with OOB

Although OOB estimation has shown good robustness and works reliably on many different datasets and models (Breiman 1996b), we still do not know whether the random variation in the number of distinct observations (which naturally comes from the classical bootstrap) actually affects its performance. Sequential Bootstrap gives us a very clean way to change only this one source of variability: it keeps the model structure, all the hyperparameters, and the way we aggregate predictions exactly the same as before. Because of this property, SB becomes a perfect tool for doing controlled experiments — we can check whether making the number of distinct observations fixed will influence accuracy-related measures or variance-related measures in OOB estimation.

Therefore, in this paper we do not present Sequential Bootstrap as a new method that is supposed to give higher accuracy. Instead, we use it as an empirical tool that helps us better understand how the details of resampling affect the behaviour of OOB-based ensemble evaluation. This approach allows us to make a direct and fair comparison between the classical OOB estimator and the Sequential-Bootstrap version (SB-OOB) under exactly the same experimental conditions.

3. Methodology

In this section we formalize the estimators studied in this work. We first introduce notation and the classical Out-of-Bag (OOB) estimator based on the multinomial bootstrap, then define the Sequential Bootstrap analogue (SB-OOB). Finally, we present a variance-structure viewpoint that motivates the empirical comparisons in Section 5.

3.1 Data and Bagging Setup

Let

$$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}.$$

denote a training sample, where $X_i \in \chi$ is a feature vector and $Y_i \in \mathcal{Y}$ is a response (class label or real-valued outcome). Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ denote a loss function, taken to be squared-error loss for regression and 0-1 loss for classification.

Bagging constructs an ensemble of base learners $\{T_b\}_{b=1}^{B}$ by fitting each T_b on a resampled version of $\mathcal{D}$. For notational convenience, let

$$I_b = (I_{b1}, \dots, I_{bN_b})$$

denote the (possibly multiset-valued) sequence of indices used to train the b-th base learner, where N_b is the (possibly random) length of the resample and each $I_{bj} \in \{1, \dots, n\}$. The corresponding training set for replicate b is

$$\mathcal{D}_b = \{(X_{I_{bj}}, Y_{I_{bj}}) : j = 1, \dots, N_b\}.$$

For an observation $i \in \{1, \dots, n\}$, define the OOB index set under a given resampling scheme as

$$\mathrm{OOB}(i) = \{b \in \{1, \dots, B\}\, i \notin \{I_{b1}, \dots, I_{bN_b}\}\},$$

that is, the set of bootstrap replicates in which observation i is not used for training.

Throughout this work, the base learners T_b are CART models fitted on $\mathcal{D}_b$ with fixed tuning parameters. The only component that differs between the classical OOB estimator and SB-OOB is the resampling mechanism generating I_b.

3.2 Classical Out-of-Bag Estimator

Under the classical bootstrap, each replicate draws exactly n indices with replacement from $\{1, \dots, n\}$. Formally, for replicate b,

$$I_{b1}, \dots, I_{bn} \overset{\text{i.i.d}}{\sim} \text{Unif}\,\{1, \dots, n\}, \quad N_b \equiv n.$$

The multiset of indices $\{I_{b1}, \dots, I_{bn}\}$ defines the training sample $\mathcal{D}_b$ for the b-th tree T_b. For each observation i, the classical OOB prediction is obtained by averaging predictions from all trees for which i is out-of-bag:

$$\hat{f}_{\text{OOB}}(X_i) = \frac{1}{|\text{OOB}(i)|} \sum_{b \in \text{OOB}(i)} T_b(X_i),$$

whenever $|\text{OOB}(i)| \geq 1$. Let

$$\mathcal{J}_{\text{OOB}} = \{i \in \{1, \dots, n\} : |\text{OOB}(i)| \geq 1\}$$

denote the set of observations that are OOB for at least one tree. The classical OOB error estimator is then

$$\widehat{\text{Err}}_{\text{OOB}} = \frac{1}{|\mathcal{J}_{\text{OOB}}|} \sum_{i \in \mathcal{J}_{\text{OOB}}} L(Y_i, \hat{f}_{\text{OOB}}(X_i)).$$

This estimator is the baseline against which we compare the Sequential Bootstrap variant.

3.3 Sequential Bootstrap and SB-OOB Estimator

Sequential Bootstrap modifies the resampling mechanism by stabilizing the number of distinct observations in each replicate rather than the total number of draws. Let $\rho \in (0, 1)$ be a target proportion, and define a target distinct-count

$$k_n = \lfloor \rho n \rfloor.$$

In this work, ρ is chosen to match the asymptotic expected fraction of distinct observations under the classical bootstrap (approximately 0.632(Efron and Tibshirani 1994)), so that

$$k_n \approx 0.632n.$$

For each replicate b, Sequential Bootstrap proceeds as follows:

1. Initialize a set of distinct indices $S_b^{(0)} = \emptyset$ and draw counter $t = 0$.
2. Repeat:

- Draw $J_{b,t+1} \sim \text{Unif}\,\{1, \dots, n\}$ independently of previous draws;
- Update $S_b^{(t+1)} = S_b^{(t)} \cup \{J_{b,t+1}\}$;
- Set $t \leftarrow t + 1$; until $|S_b^{(t)}| = k_n$.

We then define

$$I_b = (J_{b1}, \dots, J_{bT_b}), \quad N_b = T_b,$$

where T_b is the (random) stopping time at which k_n distinct indices have been collected. The corresponding training sample $\mathcal{D}_b$ is constructed as in Section 3.1.

The SB-based OOB index sets $\text{OOB}_{\text{SB}}(i)$, predictions $\hat{f}_{\text{SB-OOB}}(X_i)$, and error estimator $\widehat{\text{Err}}_{\text{SB-OOB}}$ are defined analogously to the classical case by replacing the bootstrap resamples I_b with those generated by Sequential Bootstrap and computing OOB predictions only over replicates in which observation i is not selected.

Concretely,

$$\hat{f}_{\text{SB-OOB}}(X_i) = \frac{1}{|\text{OOB}_{\text{SB}}(i)|} \sum_{b \in \text{OOB}_{\text{SB}}(i)} T_b(X_i),$$

and

$$\widehat{\text{Err}}_{\text{SB-OOB}} = \frac{1}{|\mathcal{J}_{\text{SB}}|} \sum_{i \in \mathcal{J}_{\text{SB}}} L(Y_i, \hat{f}_{\text{SB-OOB}}(X_i)),$$

where $\mathcal{I}_{\mathrm{SB}} = \{i : |\mathrm{OOB}_{\mathrm{SB}}(i)| \geq 1\}$.

The two estimators $\widehat{\mathrm{Err}}_{\mathrm{OOB}}$ and $\widehat{\mathrm{Err}}_{\text{SB-OOB}}$ therefore differ only in the resampling scheme generating their respective training replicates; all model, loss, and tuning choices are held fixed.

3.4 Variance-Structure Viewpoint

To motivate the empirical comparisons in Section 5, it is useful to view the change from classical bootstrap to Sequential Bootstrap as a modification of a particular variance component associated with resampling.

Let $\hat{\theta}_b$ denote a scalar statistic derived from the b-th replicate that enters into an OOB-based quantity of interest. Examples include node-level prediction summaries, contributions to the OOB error, or meta-prediction outputs. Let U_b denote the number of distinct training indices used in the b-th replicate, i.e.

$$U_b = |\{I_{b1}, \ldots, I_{bN_b}\}|.$$

We make no strong assumptions about the distribution of $\hat{\theta}_b$, but we note that the following identity always holds by the law of total variance.

Lemma 1 (Variance Decomposition)

For any resampling scheme and any square-integrable statistic $\hat{\theta}_b$,

$$\mathrm{Var}(\hat{\theta}_b) = \mathbb{E}[\mathrm{Var}(\hat{\theta}_b|U_b)] + \mathrm{Var}(\mathbb{E}[\hat{\theta}_b|U_b]).$$

This decomposition separates the overall variability of $\hat{\theta}_b$ into a component due to randomness conditional on the distinct-sample count and a component due to randomness in the distinct-sample count itself.

The classical multinomial bootstrap and Sequential Bootstrap can be viewed as two resampling schemes that differ primarily in the distribution of U_b. Under the classical bootstrap, U_b fluctuates around its expectation and, under standard assumptions, has variance on the order of n, whereas Sequential Bootstrap explicitly stabilizes U_b at a target value k_n. As a consequence, we expect the second term,

$$\mathrm{Var}(\mathbb{E}[\hat{\theta}_b|U_b]),$$

to be more strongly affected by the transition from classical bootstrap to SB than first term, $\mathbb{E}[\mathrm{Var}(\hat{\theta}_b|U_b)]$, which reflects variability conditional on a fixed effective sample size.

In this work we do not assume or prove that the conditional distribution of $\hat{\theta}_b$ given U_b is identical under the two resampling schemes. Instead, we use this variance decomposition as a conceptual framework: Sequential Bootstrap is interpreted as a resampling operator that stabilizes one structural aspect of the bootstrap (the distinct-sample count) while leaving model structure and loss unchanged. The empirical question, addressed in Section 5, is to what extent this stabilization manifests in observable differences between classical OOB and SB-OOB across accuracy-oriented and variance-oriented diagnostics.

4. Experimental Design

In this section we describe the experimental setup that we use to compare the classical OOB estimator with the Sequential-Bootstrap version (SB-OOB). Our design is based on a very strict controlled-perturbation idea: the only thing that is different between the two methods is the resampling scheme itself. Everything else — the model we use, all the hyperparameters, the loss function, and the way we evaluate the results — is kept exactly the same in both cases.

4.1 Datasets

We exactly follow the five groups of experiments that Breiman used in his original OOB paper (Breiman 1996b) (see Section 2.1) and use the same publicly available datasets. For the synthetic part, we take waveform, twonorm, threenorm, and ringnorm for classification problems, and Friedman 1-3 (Friedman 1991) for regression problems. For the real datasets, we use breast-cancer, diabetes, vehicle, satellite, and dna for classification tasks, and Boston and Ozone for regression tasks.

If a dataset already has an official train-test split in the original paper, we directly use that split. If there is no official split, we always do the same fixed random split: two-thirds of the data for training and one-third for testing. Importantly, we use exactly the same split for all experiments and all random seeds so that the results are fully comparable.

All datasets are used in their original form. We do not apply any additional preprocessing: no scaling, no transformation, and no feature engineering.

4.2 Learning Algorithm and Hyperparameters

All experiments employ Classification and Regression Trees (CART) as base learners, implemented with fixed hyperparameters across all datasets and all experimental runs. Specifically:

- splitting rules, stopping criteria, and pruning settings remain unchanged,
- no hyperparameter tuning procedure (grid search, cross-validation, or data-driven tuning) is used,
- the number of bagging replicates is fixed at $B = 100$ for all configurations.

The motivation for fixing CART hyperparameters is to isolate the effect of the resampling mechanism rather than confounding it with modeling decisions or tuning variability.

4.3 Resampling Protocol

For each dataset and each random seed, we construct two bagged ensembles:

1. Classical OOB Ensemble: trained using the multinomial bootstrap defined in Section 3.2.
2. SB-OOB Ensemble: trained using Sequential Bootstrap defined in Section 3.3 with the target distinct-count parameter set to approximate the classical bootstrap expectation ($\rho \approx 0.632$).

The two ensembles are matched in the following ways:

- both use the same initial random seeds for experiment replication,
- both generate exactly $B = 100$ resampling replicates,
- all model-fitting calls occur in the same order using identical computational routines and hardware,
- OOB predictions are computed using the same aggregation logic.

Thus any observed difference between the two estimators is attributable solely to the change in resampling mechanism.

4.4 Evaluation Metrics

Consistent with the original OOB study and its five experimental families (Breiman 1996b), we evaluate the following diagnostic quantities:

Differences are reported in signed form as

Table 1: Summary of diagnostic metrics evaluated across the five experimental families.

Experiment	Domain	Diagnostic.Type	Notation	Description
EXP1	Classification	Node accuracy	E1, E2	Absolute deviation between estimated & empirical class proportions
EXP2	Regression	Node accuracy	EB1, EB2	Mean squared deviation between node predictions and empirical conditional means
EXP3	Mixed	Stability	R1-R4	Within-node variability measures across bootstrap replicates
EXP4	Mixed	Generalization alignment		Absolute difference between test error and OOB error
EXP5	Regression	Meta-prediction performance	MSE	Secondary prediction using OOB-derived features

$$\Delta = \text{SB-OOB} - \text{Classical OOB},$$

to maintain consistent interpretability across all metrics.

4.5 Random Seed Replication

To make sure that our comparison is fair and to measure how stable the results are when we run the experiments many times, we repeat the whole experimental suite using three different random seeds. Importantly, for every dataset we always use exactly the same three seeds for both the classical OOB method and the SB-OOB method. In this way, any difference we see cannot come from other random factors - it can only come from the difference in the resampling mechanism itself.

Because we repeat everything with several seeds, we can not only look at the average performance difference, but also see how consistent the results are across different runs.

5. Results and Analysis

In this section we show the empirical comparison between the classical OOB estimator and the Sequential-Bootstrap version (SB-OOB). We follow exactly the same way Breiman reported his results in the original paper (Breiman 1996b): we group the experiments into EXP1–EXP5, and for each dataset we run everything under three different random seeds. All numbers we report are based on these three independent runs. To make the comparison easy to read, we mainly present the signed differences (SB-OOB minus classical OOB) together with paired summary statistics across the three seeds.

$$\Delta = \text{SB-OOB} - \text{Classical OOB}.$$

Positive values indicate higher metric values under SB-OOB.

All comparisons are made using the same data splits, base-learner configurations, and compu-

Table 2: Node-level classification probability error (E1 and E2) for OOB vs. SB-OOB across nine datasets (Seed = 1)

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	E1_B	0.02520	0.02440	-7.15e-04
waveform	synthetic	E2_B	0.02840	0.02730	-1.06e-03
twonorm	synthetic	E1_B	0.00498	0.00447	-5.04e-04
twonorm	synthetic	E2_B	0.00498	0.00447	-5.04e-04
threenorm	synthetic	E1_B	0.00552	0.00683	1.31e-03
threenorm	synthetic	E2_B	0.00552	0.00683	1.31e-03
ringnorm	synthetic	E1_B	0.03840	0.03940	9.54e-04
ringnorm	synthetic	E2_B	0.03840	0.03940	9.54e-04
breast-cancer	real	E1_B	0.00532	0.00548	1.56e-04
breast-cancer	real	E2_B	0.00532	0.00548	1.56e-04
diabetes	real	E1_B	0.00350	0.00329	-2.13e-04
diabetes	real	E2_B	0.00350	0.00329	-2.13e-04
vehicle	real	E1_B	0.00357	0.00316	-4.15e-04
vehicle	real	E2_B	0.00427	0.00376	-5.12e-04
satellite	real	E1_B	0.00177	0.00169	-8.69e-05
satellite	real	E2_B	0.00234	0.00224	-9.37e-05
dna	real	E1_B	0.00603	0.00593	-1.01e-04
dna	real	E2_B	0.00654	0.00641	-1.25e-04

tational routines. The resampling mechanism is the only experimental difference.

5.1 EXP1 - Classification Node-Level Accuracy Diagnostics

For all classification datasets, the two node-level accuracy measures (E_1, E_2) stayed extremely close between SB-OOB and classical OOB no matter which random seed we used. The differences were always very small, and we could not see any clear pattern - sometimes positive, sometimes negative. This tells us clearly that making the number of distinct samples exactly the same in every replicate does not really change the estimated class proportions inside the tree nodes for classification problems.

Overall, EXP1 results (Table 2) are consistent with the view that node-level classification accuracy is primarily determined by the tree-induction procedure and less sensitive to the composition variability induced by the classical bootstrap.

Table 3: Node-level regression error (EB1/EB2) for OOB vs. SB-OOB across five datasets (Seed = 1).

dataset	type	metric	OOB	SB_OOB	diff
friedman1	synthetic	EB1	2.100	2.110	0.005160
friedman1	synthetic	EB2	2.100	2.110	0.005160
friedman2	synthetic	EB1	241.000	241.000	0.017000
friedman2	synthetic	EB2	241.000	241.000	0.017000
friedman3	synthetic	EB1	0.154	0.155	0.000649
friedman3	synthetic	EB2	0.154	0.155	0.000649
Boston	real	EB1	5.660	5.660	-0.001580
Boston	real	EB2	5.660	5.660	-0.001580
Ozone	real	EB1	10.000	10.200	0.136000
Ozone	real	EB2	10.000	10.200	0.136000

5.2 EXP2 - Regression Node-Level Accuracy Diagnostics

For the regression datasets, the two corresponding measures (EB_1, EB_2) also showed only very small absolute differences across the three random seeds. Different from what we saw in EXP1, in EXP2 the differences were more often negative - that is, SB-OOB gave slightly smaller values than classical OOB - but the size was still quite modest. This probably means that fixing the number of distinct observations can slightly reduce the node-level deviation in regression trees, but the effect is not big enough to be considered a real or practically important improvement.

These results (Table 3) indicate that conditional-mean estimation within regression tree nodes appears robust to fluctuations in distinct sample counts under the classical bootstrap.

5.3 EXP3 - Node Stability Diagnostics

EXP3 looks at how much the predictions inside the same node vary across different bootstrap replicates. We use four node-level variability measures $(R_1 - R_4)$ to measure this. Different from EXP1 and EXP2, here we can really see some clear separation: for several of the R_j metrics, SB-OOB almost always gives smaller variability numbers than classical OOB. What is more interesting is that for some datasets, the direction of this reduction is the same no matter which random seed we use. Of course, how big the reduction is still depends on the dataset.

As we can see from Table 4, this pattern matches exactly what we said in Section 3.4 from the variance-structure point of view: when we make the number of distinct samples the same in every replicate, the tree nodes become less affected by the random change of effective sample size. So the partitioning of nodes and the final predictions become a little more stable. Although the reduction

Table 4: Within-node variability metrics (R1-R4) for OOB vs. SB-OOB across four datasets (Seed = 1).

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	R1	0.3420	0.3420	0.000000
waveform	synthetic	R2	-0.0196	-0.0101	0.009440
waveform	synthetic	R3	0.3230	0.3320	0.009360
waveform	synthetic	R4	41.7000	41.8000	0.118000
twonorm	synthetic	R1	0.0000	0.0000	0.000000
twonorm	synthetic	R2	0.0151	0.0152	0.000114
twonorm	synthetic	R3	0.0151	0.0152	0.000112
twonorm	synthetic	R4	34.7000	34.6000	-0.088700
threenorm	synthetic	R1	0.0000	0.0000	0.000000
threenorm	synthetic	R2	0.0159	0.0180	0.002140
threenorm	synthetic	R3	0.0159	0.0180	0.002140
threenorm	synthetic	R4	62.7000	62.7000	-0.044700
ringnorm	synthetic	R1	0.0000	0.0000	0.000000
ringnorm	synthetic	R2	0.6620	0.5970	-0.064700
ringnorm	synthetic	R3	0.6570	0.5930	-0.063800
ringnorm	synthetic	R4	43.1000	43.1000	0.008500

we observe is not very large, the direction is the same in almost all repeated runs.

5.4 EXP4 - OOB vs Test-Set Generalization Alignment

EXP4 checks how well the OOB error matches the true test error by looking at absolute differences and ratio-based measures. Across the three random seeds, SB-OOB and classical OOB gave almost the same numbers - the differences were always very small. Sometimes SB-OOB was a little better, sometimes a little worse, there was no clear winner. However, on several datasets we did see that SB-OOB was slightly closer to the real test error, though the improvement was quite modest.

The numbers in Table 5 support the same idea: how well OOB error matches the true test error does depend a little bit on the details of resampling, but for tree-based models the classical bootstrap already gives quite stable results. In other words, making the distinct-sample count fixed can change things in some datasets, but overall the standard bootstrap is already good enough on this point.

Table 5: OOB-test discrepancy metrics (absdiff, eOB, eTS, ratio) for OOB vs. SB-OOB (Seed = 1).

dataset	type	metric	OOB	SB_OOB	diff
twonorm	class	absdiff	0.0213	0.0187	-0.00264
twonorm	class	eOB	0.0904	0.0908	0.00040
twonorm	class	eTS	0.0816	0.0803	-0.00128
twonorm	class	ratio	1.4300	1.2600	-0.16500
friedman1	reg	absdiff	1.2500	1.2400	-0.01530
friedman1	reg	eOB	8.4200	8.4000	-0.02720
friedman1	reg	eTS	8.0100	7.9600	-0.05390
friedman1	reg	ratio	1.4000	1.3900	-0.01260

Table 6: Downstream prediction performance using OOB-based meta-features for OOB vs. SB-OOB across five datasets (Seed = 1).

dataset	type	metric	OOB	SB_OOB	diff
friedman1	reg	mse_oob_outputs	10.9000	10.9000	3.30e-02
friedman1	reg	mse_original	13.7000	13.7000	0.00e+00
friedman2	reg	mse_oob_outputs	27248.0000	26844.0000	-4.04e+02
friedman2	reg	mse_original	38010.0000	38010.0000	0.00e+00
friedman3	reg	mse_oob_outputs	0.0405	0.0404	-1.31e-04
friedman3	reg	mse_original	0.0439	0.0439	0.00e+00
Boston	reg	mse_oob_outputs	19.9000	19.7000	-1.90e-01
Boston	reg	mse_original	20.2000	20.2000	0.00e+00
Ozone	reg	mse_oob_outputs	5755.0000	5833.0000	7.85e+01
Ozone	reg	mse_original	8790.0000	8790.0000	0.00e+00

5.5 EXP5 - Meta-Prediction Diagnostics

EXP5 takes the OOB predictions from all trees as new features and trains a second-level model on them. Then we look at the mean-squared-error of this meta-model as our diagnostic. In many dataset-seed combinations, SB-OOB gave slightly lower values than classical OOB. This tells us that when we use OOB quantities not directly as the final answer, but as input features for another learner, stabilizing the distinct-sample count might bring some small stability benefit downstream.

From Table 6 we can see that the difference is still quite small, but this experiment clearly shows one thing: even if we do not see big changes when we use OOB error directly, fixing the number of distinct observations can still matter in situations where OOB-derived numbers are used indirectly.

5.6 Cross-Seed and Cross-Experiment Summary

From all five groups of experiments, we can see a very clear and consistent pattern:

1. For accuracy-oriented measures (EXP1 and EXP2), SB-OOB and classical OOB gave almost exactly the same results - basically no change at all.
2. For stability and variability-oriented measures (EXP3, EXP4, and EXP5), SB-OOB usually gave slightly smaller numbers. For quite a few metrics and datasets, the direction of this reduction was the same no matter which random seed we used. Of course, there were still some differences between datasets.
3. All the differences we observed were quite small. This tells us that the classical bootstrap is already very robust when we use it for OOB evaluation in tree-based ensembles.

Putting everything together, our results show that Sequential Bootstrap does not change the predictive accuracy of OOB estimation, but it can work as a useful stability-focused tool. It helps us see some parts of the variance structure of OOB that are hidden when we only use the standard bootstrap.

6. Discussion

In this work we carefully studied how making the number of distinct observations exactly the same in every bootstrap replicate affects Out-of-Bag (OOB) estimation. We did this by replacing the classical bootstrap with Sequential Bootstrap (SB) while keeping everything else completely unchanged - the tree model, all hyperparameters, the way we aggregate predictions, everything. Because of this strict design, any difference we see can only come from the randomness in the distinct-sample count itself. This lets us clearly check whether this particular source of variability

really matters for various OOB-based diagnostic measures.

After running all five experimental families on both synthetic and real datasets and repeating everything three times with different random seeds, we found two very clear main results.

First, all accuracy-oriented measures - no matter classification or regression - stayed almost exactly the same when we switched from classical bootstrap to SB. This tells us that the standard bootstrap already gives very stable OOB predictions even though the effective sample size changes randomly from tree to tree.

Second, several variance-oriented and stability-oriented measures became slightly smaller under SB-OOB. What is more important is that for many of them the reduction went in the same direction across all three seeds, especially for node-level variability and for the meta-prediction experiment. These patterns match perfectly the variance-structure viewpoint we explained in Section 3.4 (although of course we are not claiming any strict mathematical inequality or a general theorem here).

So, the best way to understand our results is that they mainly tell us how robust the classical OOB estimator really is, rather than showing that SB-OOB is a better method that can improve predictive performance. In particular, our experiments clearly confirm that when we only care about accuracy, classical OOB is already very insensitive to the random change in distinct-sample count. At the same time, they also suggest that Sequential Bootstrap can work as a very useful stability-focused perturbation tool - it helps us look deeper into the internal variance structure that is usually hidden under the standard bootstrap.

From the methodology point of view, this work gives us a much clearer picture of how the details of the resampling step affect the behaviour of ensemble evaluation. But it definitely does not mean we need to throw away the classical bootstrap in real applications.

There are still several limitations that we should point out, and also many interesting directions for future work.

First, our study is purely empirical. We did not try to give any theoretical analysis of the SB-OOB estimator. It would be very valuable if someone could derive exact conditions under which stabilizing the distinct-count actually reduces variance.

Second, we only used simple CART trees with fixed hyperparameters. The effect might be different (maybe larger) for deeper random forests, extremely randomized trees, or other highly unstable high-capacity learners.

Finally, we always set the target distinct-count to the same expected value as the classical bootstrap ($\rho \approx 0.632$). It would be quite interesting to see what happens if we choose other targets - for example $0.5n$ or $0.9n$ - and whether there are systematic trade-offs.

Future work can therefore go in several directions:

(i) test the same idea on more modern ensemble methods and much larger model sizes,
(ii) combine distinct-count stabilization with other sources of randomness (like random feature selection or split-point perturbation),
(iii) develop theoretical results for SB-OOB under simplified settings,
(iv) check whether SB can be used as a diagnostic tool or even as a mild regularization trick in complicated multi-stage learning systems.

7. Conclusion

In summary, we carefully studied how making the number of distinct observations exactly the same in every bootstrap replicate affects Out-of-Bag (OOB) estimation. We did this by building the Sequential-Bootstrap version (SB-OOB) and comparing it with the classical version under a very strict controlled setting. We exactly reproduced Breiman's five original experimental families, and the only thing we changed was the resampling method - all the tree models, all hyperparameters, the number of trees, everything else stayed completely the same. Because of this design, any difference we see can only come from the random change in the number of distinct samples.

Our results show very clearly that accuracy-related measures (both classification and regression) stayed almost exactly the same when we switched to Sequential Bootstrap. But for several variance-sensitive and stability-sensitive measures, SB-OOB gave slightly smaller values, and the direction was usually the same across different random seeds. This tells us that the classical OOB estimator is already quite robust to the random fluctuation in distinct-sample count when we only care about prediction accuracy. At the same time, Sequential Bootstrap can work as a very clean diagnostic tool that helps us see the variance-structure part that is usually hidden under the standard bootstrap - it is not a method to make the model predict better.

In short, this work gives us a lot of solid empirical evidence that helps us understand much better how the details of the resampling step affect OOB evaluation in tree-based ensembles. Our results clearly show that there is no need at all to replace the classical bootstrap in real applications. What we do get is a clearer picture of how stable the classical OOB estimator really is, and also some new ideas that can inspire future theoretical work or experiments with other kinds of models and other resampling methods.

References

Babu, G Jogesh, PK Pathak, and CR Rao. 1999. "Second-Order Correctness of the Poisson Bootstrap." *The Annals of Statistics* 27 (5): 1666–83.

———. 2000. "Consistency and Accuracy of the Sequential Bootstrap." In *Statistics for the 21st Century*, 37–48. CRC Press.

Barbe, Philippe, and Patrice Bertail. 2012. *The Weighted Bootstrap*. Vol. 98. Springer Science & Business Media.

Bickel, Peter J, and David A Freedman. 1984. "Asymptotic Normality and the Bootstrap in Stratified Sampling." *The Annals of Statistics*, 470–82.

Breiman, Leo. 1996a. "Bagging Predictors." *Machine Learning* 24 (2): 123–40.

———. 1996b. "Out-of-Bag Estimation."

———. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman; Hall/CRC.

Friedman, Jerome H. 1991. "Multivariate Adaptive Regression Splines." *The Annals of Statistics* 19 (1): 1–67.

Ho, Tin Kam. 1998. "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832–44.

Janitza, Silke, and Roman Hornung. 2018. "On the Overestimation of Random Forest's Out-of-Bag Error." *PloS One* 13 (8): e0201904.

Politis, Dimitris N, and Joseph P Romano. 1994. "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions." *The Annals of Statistics*, 2031–50.

Rao, C Radhakrishna, PK Pathak, and VI Koltchinskii. 1997. "Bootstrap by Sequential Resampling." *Journal of Statistical Planning and Inference* 64 (2): 257–81.

Table 7: A.1 EXP1 Seed25

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	E1_B	0.02400	0.02390	-1.08e-04
waveform	synthetic	E2_B	0.02650	0.02640	-1.61e-04
twonorm	synthetic	E1_B	0.00618	0.00634	1.61e-04
twonorm	synthetic	E2_B	0.00618	0.00634	1.61e-04
threenorm	synthetic	E1_B	0.00739	0.00619	-1.20e-03
threenorm	synthetic	E2_B	0.00739	0.00619	-1.20e-03
ringnorm	synthetic	E1_B	0.03730	0.03630	-1.04e-03
ringnorm	synthetic	E2_B	0.03730	0.03630	-1.04e-03
breast-cancer	real	E1_B	0.00497	0.00566	6.82e-04
breast-cancer	real	E2_B	0.00497	0.00566	6.82e-04
diabetes	real	E1_B	0.00351	0.00358	6.77e-05
diabetes	real	E2_B	0.00351	0.00358	6.77e-05
vehicle	real	E1_B	0.00315	0.00327	1.16e-04
vehicle	real	E2_B	0.00377	0.00386	9.02e-05
satellite	real	E1_B	0.00173	0.00170	-2.28e-05
satellite	real	E2_B	0.00228	0.00228	1.30e-06
dna	real	E1_B	0.00617	0.00606	-1.04e-04
dna	real	E2_B	0.00668	0.00658	-1.03e-04

Appendix A: Full EXP1-EXP5 Tables

In the main text, we report detailed numerical tables for a representative seed (seed = 1). For completeness and reproducibility, full results for additional seeds (25 and 50) are presented in Appendix A. The qualitative conclusions remain unchanged across seeds.

Table 8: A.2 EXP2 Seed25

dataset	type	metric	OOB	SB_OOB	diff
friedman1	synthetic	EB1	2.090	2.100	0.012400
friedman1	synthetic	EB2	2.090	2.100	0.012400
friedman2	synthetic	EB1	242.000	243.000	0.188000
friedman2	synthetic	EB2	242.000	243.000	0.188000
friedman3	synthetic	EB1	0.155	0.156	0.000712
friedman3	synthetic	EB2	0.155	0.156	0.000712
Boston	real	EB1	5.580	5.590	0.007270
Boston	real	EB2	5.580	5.590	0.007270
Ozone	real	EB1	8.070	8.280	0.217000
Ozone	real	EB2	8.070	8.280	0.217000

Table 9: A.3 EXP3 Seed25

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	R1	0.38700	0.38700	0.00000
waveform	synthetic	R2	0.00545	-0.00589	-0.01130
waveform	synthetic	R3	0.39200	0.38100	-0.01120
waveform	synthetic	R4	42.10000	42.00000	-0.09020
twonorm	synthetic	R1	0.00000	0.00000	0.00000
twonorm	synthetic	R2	0.01590	0.02020	0.00427
twonorm	synthetic	R3	0.01590	0.02020	0.00427
twonorm	synthetic	R4	34.20000	34.10000	-0.05610
threenorm	synthetic	R1	0.00000	0.00000	0.00000
threenorm	synthetic	R2	0.01800	0.02940	0.01140
threenorm	synthetic	R3	0.01800	0.02930	0.01140
threenorm	synthetic	R4	63.00000	62.90000	-0.04420
ringnorm	synthetic	R1	0.00000	0.00000	0.00000
ringnorm	synthetic	R2	0.69100	0.60700	-0.08390
ringnorm	synthetic	R3	0.68500	0.60200	-0.08270
ringnorm	synthetic	R4	43.10000	43.00000	-0.06140

Table 10: A.4 EXP4 Seed25

dataset	type	metric	OOB	SB_OOB	diff
twonorm	class	absdiff	0.0201	0.0212	0.001110
twonorm	class	eOB	0.0882	0.0923	0.004100
twonorm	class	eTS	0.0842	0.0840	-0.000152
twonorm	class	ratio	1.2700	1.3500	0.084500
friedman1	reg	absdiff	1.3700	1.3600	-0.003010
friedman1	reg	eOB	8.5500	8.5200	-0.036400
friedman1	reg	eTS	8.5000	8.5000	0.004170
friedman1	reg	ratio	1.4300	1.4200	-0.006260

Table 11: A.5 EXP5 Seed25

dataset	type	metric	OOB	SB_OOB	diff
friedman1	reg	mse_oob_outputs	10.8000	10.9000	6.44e-02
friedman1	reg	mse_original	13.2000	13.2000	0.00e+00
friedman2	reg	mse_oob_outputs	27029.0000	26671.0000	-3.58e+02
friedman2	reg	mse_original	38416.0000	38416.0000	0.00e+00
friedman3	reg	mse_oob_outputs	0.0411	0.0415	3.76e-04
friedman3	reg	mse_original	0.0437	0.0437	0.00e+00
Boston	reg	mse_oob_outputs	18.0000	18.8000	7.53e-01
Boston	reg	mse_original	22.1000	22.1000	0.00e+00
Ozone	reg	mse_oob_outputs	5690.0000	5531.0000	-1.59e+02
Ozone	reg	mse_original	9178.0000	9178.0000	0.00e+00

Table 12: A.6 EXP1 Seed50

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	E1_B	0.02250	0.02250	-2.39e-05
waveform	synthetic	E2_B	0.02520	0.02530	4.49e-05
twonorm	synthetic	E1_B	0.00507	0.00569	6.12e-04
twonorm	synthetic	E2_B	0.00507	0.00569	6.12e-04
threenorm	synthetic	E1_B	0.00507	0.00764	2.57e-03
threenorm	synthetic	E2_B	0.00507	0.00764	2.57e-03
ringnorm	synthetic	E1_B	0.03830	0.03890	5.89e-04
ringnorm	synthetic	E2_B	0.03830	0.03890	5.89e-04
breast-cancer	real	E1_B	0.00468	0.00473	4.90e-05
breast-cancer	real	E2_B	0.00468	0.00473	4.90e-05
diabetes	real	E1_B	0.00367	0.00360	-7.86e-05
diabetes	real	E2_B	0.00367	0.00360	-7.86e-05
vehicle	real	E1_B	0.00305	0.00319	1.41e-04
vehicle	real	E2_B	0.00363	0.00378	1.54e-04
satellite	real	E1_B	0.00169	0.00175	5.68e-05
satellite	real	E2_B	0.00225	0.00234	9.05e-05
dna	real	E1_B	0.00583	0.00589	6.38e-05
dna	real	E2_B	0.00637	0.00637	2.50e-06

Table 13: A.7 EXP2 Seed50

dataset	type	metric	OOB	SB_OOB	diff
friedman1	synthetic	EB1	2.130	2.120	-0.013400
friedman1	synthetic	EB2	2.130	2.120	-0.013400
friedman2	synthetic	EB1	246.000	246.000	-0.075200
friedman2	synthetic	EB2	246.000	246.000	-0.075200
friedman3	synthetic	EB1	0.154	0.153	-0.000543
friedman3	synthetic	EB2	0.154	0.153	-0.000543
Boston	real	EB1	5.870	5.880	0.008830
Boston	real	EB2	5.870	5.880	0.008830
Ozone	real	EB1	7.720	8.080	0.363000
Ozone	real	EB2	7.720	8.080	0.363000

Table 14: A.8 EXP3 Seed50

dataset	type	metric	OOB	SB_OOB	diff
waveform	synthetic	R1	0.3390	0.3390	0.00000
waveform	synthetic	R2	-0.0204	0.0049	0.02530
waveform	synthetic	R3	0.3180	0.3440	0.02520
waveform	synthetic	R4	42.0000	42.0000	0.01410
twonorm	synthetic	R1	0.0000	0.0000	0.00000
twonorm	synthetic	R2	0.0180	0.0147	-0.00328
twonorm	synthetic	R3	0.0180	0.0147	-0.00328
twonorm	synthetic	R4	34.2000	34.2000	-0.01050
threenorm	synthetic	R1	0.0000	0.0000	0.00000
threenorm	synthetic	R2	0.0254	0.0323	0.00688
threenorm	synthetic	R3	0.0254	0.0323	0.00686
threenorm	synthetic	R4	62.9000	63.0000	0.04320
ringnorm	synthetic	R1	0.0000	0.0000	0.00000
ringnorm	synthetic	R2	0.5770	0.5850	0.00781
ringnorm	synthetic	R3	0.5730	0.5810	0.00781
ringnorm	synthetic	R4	43.8000	43.7000	-0.12300

Table 15: A.9 EXP4 Seed50

dataset	type	metric	OOB	SB_OOB	diff
twonorm	class	absdiff	0.0256	0.0254	-0.000196
twonorm	class	eOB	0.0891	0.0911	0.002000
twonorm	class	eTS	0.0864	0.0861	-0.000268
twonorm	class	ratio	1.6000	1.5900	-0.012500
friedman1	reg	absdiff	1.1000	1.1800	0.081900
friedman1	reg	eOB	8.2600	8.1200	-0.146000
friedman1	reg	eTS	8.2500	8.2400	-0.008680
friedman1	reg	ratio	1.1800	1.2700	0.086800

Table 16: A.10 EXP5 Seed50

dataset	type	metric	OOB	SB_OOB	diff
friedman1	reg	mse_oob_outputs	10.6000	10.7000	2.52e-02
friedman1	reg	mse_original	13.3000	13.3000	0.00e+00
friedman2	reg	mse_oob_outputs	27106.0000	27352.0000	2.47e+02
friedman2	reg	mse_original	38311.0000	38311.0000	0.00e+00
friedman3	reg	mse_oob_outputs	0.0412	0.0400	-1.23e-03
friedman3	reg	mse_original	0.0435	0.0435	0.00e+00
Boston	reg	mse_oob_outputs	19.4000	19.4000	-5.02e-02
Boston	reg	mse_original	21.1000	21.1000	0.00e+00
Ozone	reg	mse_oob_outputs	6236.0000	6066.0000	-1.70e+02
Ozone	reg	mse_original	9727.0000	9727.0000	0.00e+00