Where to Experiment?
A Best Subsets Ensemble for Purposive Site Selection

Adam Bouyamourn*

July 17, 2024

Abstract

Choosing where to conduct an experiment when a researcher is faced with a universe of possible sites is an important active area of research. This is a well-defined discrete optimization problem, in which an experimenter must make a discrete choice when faced with a larger set of possible sites. We propose a novel machine learning method for selecting experimental sites when either the Population Average Treatment Effect or the Conditional Average Treatment Effect is of interest. Recent advances in Mixed Integer Optimization have dramatically improved the practical utility of the Best Subset selection method in machine learning. We develop an ensemble method using Best Subsets as a base learner, and use it to solve a novel representation of the site selection problem as a weighted-quantile discrepancy minimization problem. We show that this method has superior average-case performance compared to alternative methods across a set of naturalistic simulation studies.

*PhD Candidate, Charles and Louise Travers Department of Political Science, University of California, Berkeley

1 Introduction

An intuitive approach to site selection is 'double randomization', in which a set of sites is randomly selected from a target population, before a randomized experiment is conducted in each site. While randomization benefits from strong asymptotic guarantees that allow for valid causal inference, 'double randomization' for multi-site experimental designs has significant drawbacks compared to a two-stage procedure in which sites are first purposively selected, before a randomized experiment is conducted in each site.

The intuition is that randomization can be inefficient for estimating a population quantity of interest, because it can assign non-zero probability mass to treatment assignment profiles that do not well-represent the quantity that we are interested in. Randomization can ensure that samples are balanced in expectation, but this property does not guarantee representativeness of any given sample drawn by randomization. In practice, we might be able to do better by purposively selecting sites, then randomizing conditional on a site selection (Kasy, 2016; Kallus, 2018). Further, the choice of sites itself may be a nuisance parameter: since we are running an experiment conditional on site selections, we can be unconcerned about the effects of choosing sites on the internal validity, except to the extent that they allow us to extrapolate. Optimizing non-randomly at the site selection level does not harm the causal inference guarantees we get from running an experiment at each chosen site.

The problem may be particularly acute when selecting small numbers of sites from a diverse population. Adapting a result of Li and Ding (2016), the variance of a randomized site selection can be given as follows. Let $\{\tau_s, \ldots, \tau_P\}$ be a finite population of sites, from which we draw a random sample of size $S < N$. The variance of the sample $\bar{\tau_S}$ under random sampling can be shown to be:

$$Var(\tau_S) = \left(\frac{1}{S} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{s=1}^{P} (\tau_s - \bar{\tau}_N)^2 \tag{1}$$

Under regularity assumptions, this quantity is asymptotically normal. Convergence to the true population variance may be guaranteed by randomization, but there is no guarantee that any particular randomized assignment of site selections will be representative of a larger population. In situations when we are considering a small number of possible experimental sites sampled from a larger population, where $S << N$, we might be able to do better by purposive selection. The alternative would be to minimize this quantity directly, by choosing a subset S that purposively minimizes the mean squared error of the

decision:

$$\min_{S} MSE[\tau - \tau(S)] = \sum_{s \in S} (\tau(S) - \bar{\tau}_N)^2 \qquad \text{s.t. } ||S||_0 \leq K \tag{2}$$

To solve problems of this kind, we present a novel ensemble learning method that benefits from recent developments in Mixed Integer Optimization that have led to significant performance improvements for the classical Best Subsets method in machine learning (Bertsimas et al., 2015). Best Subsets finds K-sparse solutions to estimation problems, which makes it especially appropriate for site selection problems in which a Researcher faces a budget constraint, in which they can experiment in exactly K sites. Further, K-sparse solutions to ℓ_0 minimization problems can be shown to have better worst-case performance than approximate solutions based on ℓ_1 minimization (for instance, the LASSO) (Zhang et al., 2014). While Best Subsets has historically been less popular than variants of LASSO and Ridge regression for regularized estimation problems, this has mostly been because, until recently, algorithms to implement Best Subsets have either been computationally infeasible for medium-sized problem instances, or do not find optimal solutions. With the advent of new tools from implementing Best Subsets, it is now possible to use it for large-sized problem instances. Further, this improved performance makes it possible to develop ensemble methods based on Best Subsets, in which the base algorithm is implemented many times with updating based on empirical losses (Zhou, 2012; Littlestone and Warmuth, 1994).

We show that our method has favourable empirical performance to rival methods, in terms of both in-sample performance and out-of-sample performance, in a naturalistic simulation study based on Auerbach and Thachil (2018). Because this study involves a large universe of sites, it is possible to assess both the internal and external validity of our approach by evaluating the sites chosen by our and rival methods on both an 'observed' and 'unobserved' set of sites. This gives us insight into the finite sample performance of our method, and its performance in a context in which realistic distribution shift is induced by differences between unobserved and observed sets.

Our method allows Researchers to combine information from pilot studies, where they have been undertaken, to improve the quality of the site selection. Without additional information, our approach seeks to minimize the observed Wasserstein distance, a metric used in the optimal transport literature, between sites included and not included sites. However, with some knowledge of the outcome model, from a pilot study, we can weight the information in the covariates, so that more prognostic covariates are

given more weight in the site selection procedure. This can close the gap between observed and oracle performance.

Our method also allows Researchers to incorporate welfare weighting in their analysis. We use a Weighted Quantile Discrepancy metric to measure the distance between distributions (Fan et al., 2022).[1] While, by default, quantiles are not weighted, to ensure that we choose sites that minimize an approximation to the ECDF, Researchers can incorporate welfare-weights and prior knowledge to ensure that the resulting selection places greater weight on sites that better represent the bottom α-% of units, for instance.

Our method is applicable to a number of other practical problems in statistics and Political Science. First, it can be used to select units to enroll in an experiment. Second, it can be used to assign treatment assignments to a fixed set of individuals. Finally, we can also use this method for cluster randomization, when a Researcher seeks to assign treatment at the cluster-level, once sites have already been selected.

To our knowledge, our paper is the first to propose an ensemble method based on Best Subsets. It is also the first to use weighted-quantile discrepancy minimization as a metric in the site selection problem.

Several notes of caution are required. The first is that there may be no one 'best' site selection to reduce generalization error from one context to another. Especially under distribution shift, it is always possible that the population differs from the observed distribution of covariates in ways that cannot be accounted for ex ante. A follow-up line of research is to incorporate insights from Distributionally-Robust Optimization into site selection (Duchi and Namkoong, 2017, 2020). These methods consider optimization procedures that improve worst-case performance when the true distribution is 'close' to the observed distribution, in terms of f-divergence, Wasserstein distance, or another discrepancy metric. Such methods offer improved worst-case performance, but at some loss to average case performance, and under assumptions about the nature of the distribution shift. The second is that prognostic covariates are necessary, and that site selection optimization procedures cannot work well in the presence of unmeasured confounders (Bicalho et al., 2022). Any site selection method based on discrepancy minimization requires that covariates are relevant to the outcome. There is no alternative to good data collection: an automated site selection tool cannot help you if you do not collect good covariate data ex ante.

[1] This is related to minimizing the observed 1-Wasserstein distance, which can be expressed as the distance between two CDFs.

1.1 Literature Review

Discrepancy minimization, in math and computer science (Chazelle, 2000; Beck and Fiala, 1981; Bansal et al., 2022), studies the problem of finding a 'colouring' (set of inclusion indicators) of a set that minimizes an error function between the sets that are assigned a given colour and the sets that are not. A set of fundamental results in this literature provide bounds on the efficiency of procedures for selecting colourings of sets that minimize the discrepancy (maximize the representativeness) between coloured and uncoloured sets.

These results have been applied to the problem of *covariate balancing*, in which a researcher aims to find a set of vectors that minimize observed discrepancy between covariates in the treatment group and covariates in the control group. Harshaw et al. (2023) apply the Gram-Schmidt Walk, a method for solving the discrepancy minimization problem developed by Bansal et al. (2017), to the problem of finding a covariate balancing vector without randomization.

Several approaches to developing site selection methods have been proposed in the literature in statistics, computer science, and political science.

Addanki et al. (2022) use the Gram-Schmidt Walk to both 1) select units for inclusion into an experiment and 2) find a treatment assignment vector that optimally balances assignments across units. This approach benefits from the statistical guarantees of the Gram-Schmidt process for solving discrepancy minimization problems, but does not produce vectors that select an exact number of sites for inclusion in an experiment.

Egami and Lee (2024) develop a method for site selection based on the Synthetic Control Method (Alberto Abadie and Hainmueller, 2010; Abadie et al., 2015; Xu, 2017; Sun et al., 2023). They choose sites that, up to weights, well-approximate sites that are not included in a given selection. This approach emphasizes choosing sites that create a large convex hull, within which excluded sites are contained.

Tipton (2013) proposes a method that uses cluster analysis to identify strata of sites. Given those strata, sites are then randomly sampled from each stratum.

Gechter et al. (2024) develops a Bayesian method, in which a prior distribution over treatment effects based on pilot data, and a prior welfare function for aggregating estimated treatment effects, are assumed. This enables the Researcher to use Monte Carlo simulations to evaluate the expected welfare under different site selections: this simulated posterior then generates estimates of the best (from the perspective of ex

post welfare) subset of sites in which to experiment.

We also contribute to a literature on ensemble methods in political science. Montgomery et al. (2012) introduced ensemble methods to political science by motivating Ensemble Bayesian Model Averaging. Grimmer et al. (2017, 2021) provide overview of ensemble methods in political science, and Samii et al. (2016) applies the procedure to analysing the results of an experiment to reduce recidivism in Colombia More recently, ensemble methods have been used across causal inference to learn heterogeneous treatment effects, with Random Forests, the SuperLearner approach, and the X-Learner being three recent examples of popular methodologies used by social scientists to learn heterogeneous causal effects (Wager and Athey, 2018; Athey et al., 2019; van der Laan et al., 2007; van der Laan and Petersen, 2007; van der Laan and Rose, 2011; Künzel et al., 2019).

Our paper is relevant to an important literature in External Validity, generalization, and distribution shift. External Validity has emerged as a key concern in empirical research in the social sciences following important an critique from (Deaton and Cartwright, 2018). Shadish et al. (2002) describe representation (of a particular population) and generalization (to a particular target population) as the key goals of external validity. Slough and Tyson (2023) develop a formal framework for external validity based on Blackwell experiments, which offers an information-geometric formulation of the notion of a space of experiments (Murray and Rice, 1993). Within their framework, we consider the problem of minimizing target discrepancy between a set of settings S and a universe of settings P. Findley et al. (2016) and Hartman (2021) provide formal frameworks for thinking about external validity in applied research contexts. Our goal is to consider the set of sites that are C-valid based on observables. Like Hartman (2021), we require a contextual exclusion restriction, or the assumption of no unobserved moderators, in order successfully choose sites.

Our approach is also related conceptually to response surface methods and optimal experimental design in the response surface literature (Box and Draper, 1987,?). Our covariate information represents an prior response surface over the set of sites, and our goal is to choose a set of sites with a similar response surface to that of the universe of observed sites. With pilot data about the response surface, we can choose a set of sites that minimizes the posterior divergence between a set of sites and the universe of observed sites, given the pilot data. We consider this proposal in an extension.

2 Method: A Best Subsets Ensemble for Site Selection

2.1 Problem of Site Selection

A Researcher observes a universe of potential experimental sites, which we denote $\mathcal{P} = \{s_1, \ldots, s_P\}$. These sites constitute the population. We suppose that the Researcher observes site-level characteristics across multiple covariates for each potential site in the study. That is, we observe a data matrix $\boldsymbol{X} = \{X_{js}\}_{j=1,s=1}^{J,P}$, where J is the dimension of covariates and P is the number of sites in the population.

The Researcher must choose a proper subset $S \subset \mathcal{P}$ of sites, subject to an ℓ_0-norm, or *cardinality constraint* on S.[2] That is, we have that, for some $0 < K < ||P||_0$:

$$||S||_0 = \sum_{s=1}^{||P||_0} \mathbb{I}\{s_s \in S\} \leq K \tag{3}$$

We suppose that the Researcher is interested in either the Population Average Treatment Effect (PATE) or the Conditional Average Treatment Effect (CATE) over the full population:

Definition 1 (Causal Estimands)

$$PATE = \mathbb{E}_{i\sim\mathcal{P}}[Y_i(1) - Y_i(0)] \qquad\qquad CATE = \mathbb{E}_{i\sim\mathcal{P}}[Y_i(1) - Y_i(0)|X = x]$$

Once the experiment is conducted, the Researcher will observe only the sample analogues to these quantities. But *which* sample analogue the Researcher will observe will depend on which sites are chosen for experimentation.

Let $\mathcal{L} : \boldsymbol{S} \times \mathcal{X} \to \mathbb{R}$ a loss function. Then, the researcher's problem is to solve:

Problem 1 (Infeasible Site Selection Problem)

$$\min_{\boldsymbol{S}} \quad \{\mathbb{E}[\mathcal{L}(PATE - SATE(\boldsymbol{S}))]\} \quad s.t.\ ||\boldsymbol{S}||_0 \leq K$$

Problem 1 is *infeasible*, for the simple reason that, since the experiment has not yet been conducted, *neither* potential outcome can be observed.[3]

[2] $||\cdot||_0$ is the ℓ_0 norm, so that $||\boldsymbol{S}||_0 = \#\{s : s \in \boldsymbol{S}\} = \sum_{s\in\mathcal{P}} \mathbb{I}\{s_s \in S\}$

[3] After Holland (1986), we might call this the Fundamental Problem of Not Having Gathered Any Outcome Data Yet.

2.2 Using Covariates As Proxies for Potential Outcomes

We can, however, use covariate information as proxies for potential outcomes. In order to do this effectively, we need to assume that covariate information is at least somewhat prognostic for potential outcomes, and that it is unconfounded: that there are no interactions between observed covariates and unobserved confounders.

Assumption 1 (No unobserved moderators)

$$\forall i, U : Y_i(Z = z, X = x, U = u) = Y_i(Z = z, X = x)$$

$$\forall i, U : Y_i(1) - Y_i(0) | X \perp\!\!\!\perp U$$

Since, by Assumption 1, potential outcomes do not depend on unobserved confounders, we can write treatment effects as an additive function of observed covariates, so that $\tau_i = g(X_i) + \epsilon_i$ (Imai et al., 2008). This assumption is also known as the Contextual Exclusion Restriction (Egami and Hartman, 2023). In **??**, we consider a partial relaxation of this assumption, in which we allow treatment effects

This allows us to perform the following decomposition of the error due to site selection:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(PATE - SATE(\boldsymbol{S}))] &= \mathbb{E}_{i\sim\mathcal{P}}[Y_i(1) - Y_i(0)] - \mathbb{E}_{i\sim\boldsymbol{S}}[Y_i(1) - Y_i(0)] && (4)\\
&= \mathbb{E}_{i\sim\mathcal{P}}[g(X_i) + \epsilon_i] - \mathbb{E}_{i\sim\boldsymbol{S}}[g(X_i) + \epsilon_i] && (5)\\
&= \int_i g(X_i | i \in P) - \int_i g(X_i | i \in \boldsymbol{S}) && (6)\\
&= \int g(X)\, d[f^P(X) - f^{\boldsymbol{S}}(X)] && (7)
\end{aligned}$$

Since $g(X)$ is unknown, we replace it with the identity function.[4] This gives us the following *discrepancy minimization problem* (Banaszczyk, 1998; Chazelle, 2000; Levy et al., 2017; Bansal et al., 2022):

Problem 2 (Discrepancy minimization for the PATE)

$$\min_{\boldsymbol{S}} \int X\, d[f^P(X) - f^{\boldsymbol{S}}(X)] \qquad s.t.\ ||\boldsymbol{S}||_0 \leq K$$

[4] In section 4.1, we consider the case where an estimate $\hat{g}(X)$ is available from a pilot study. In general, this will improve the performance of the method.

Likewise, from Hill (2011), we take the following definition of error in estimating individual treatment effects (see also Ding et al. (2017); Tipton and Mamakos (2023)):

Definition 2 (Precision in Estimation of Heterogeneous Effect (PEHE))

$$\epsilon_{PEHE} = \int_X (\hat{\tau}(x) - \tau(x))^2 p(x)\, dx$$

The Researcher wishes to minimize, by choice of $\boldsymbol{S}$,

$$\min_{\boldsymbol{S}} \quad \epsilon_{PEHE}(\boldsymbol{S}) = \int_X (\hat{\tau}(x) - \tau(x))^2 f^{\boldsymbol{S}}(x) + \int_X (\hat{\tau}(x) - \tau(x))^2 f^P(x) - f^{\boldsymbol{S}}(x) \tag{8}$$

The first term in each integral is unobserved, as before, which means that the minimization problem becomes:

Problem 3 (Discrepancy minimization for the CATE)

$$\min_{\boldsymbol{S}} \int_X f^P(x) - f^{\boldsymbol{S}}(x) \qquad s.t.\ ||\boldsymbol{S}||_0 \leq K$$

2.3 Best Subsets: A Brief Overview

The classical best subset selection problem in statistics and machine learning is the problem of choosing some finite subset of predictors that best explain the outcome in a regression model (Hocking and Leslie, 1967; Beale et al., 1967; Breiman, 1995; Miller, 2002; James et al., 2021; Thompson, 2022). This is modelled as the optimization problem:

$$\min_{\beta} \quad \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2 \qquad \text{s.t.} \quad ||\beta||_0 \leq K \tag{9}$$

Where J is the dimension of the covariates, K is an integer in $(0, \min\{n-1, J\})$. Best subsets has the attractive feature of imposing a K-*sparse* solution – that is, one in which exactly K coefficients are non-zero – on the regression problem.

Until recently, however, provably optimal solutions to this problem in the $N < P$ regime were generally computationally infeasible. Best subset selection is nonconvex and NP-hard (Natarajan, 1995). This is because any exact solution by brute force search involves $O\left(\binom{P}{K}\right)$. The implementation `leaps`, based on

a branch-and-bound technique, does not work for $p \geq 30$ (Furnival and Wilson, 1974). Greedy methods based on heuristics, such as stepwise search, are not guaranteed to find optimal solutions.

Recent work by Bertsimas et al. (2015) uses advances in Mixed Integer Optimization to develop an algorithm to implement best subsets with provably optimal solutions in the $N < P$ regime, and short runtimes.[5] The empirical performance of this method is studied by Hastie et al. (2020), who find that it generally outperforms the lasso in high signal-to-noise regimes. Two packages to implement best subsets using MIO solvers in R, `Best Subsets` and `rss` have been developed by Hastie et al. (2020) and Thompson (2022).

2.4 Using Best Subsets to Select Sites

In order to use Best Subsets to solve Problem 2 and Problem 3, our goal is to minimize the observed discrepancy between a subset of sites and the full population of sites.Whereas Best Subsets is usually applied to a data structure that is $N \times J$, we here apply it to a data structure that is $J \times S$. That is, we treat sites as *predictors*, and statistics of the covariates as our outcome. The goal is to pick the K sites that best predict features of the covariate distribution.

Given a K-sparse solution to this site selection problem, it is straightforward to extract the set of sites that 'best' predicted statistics of the covariates. These are the sites that solve the discrepancy minimization problems described above.

2.4.1 Site Selection for the PATE

To solve Problem 2, we want to minimize the discrepancy at the average. A simple argument shows that minimizing the ℓ_2 norm also minimizes the KL divergence between two distributions; Best Subsets then finds the K-sparse solution that minimizes the KL divergence between f^P and $f^{\boldsymbol{S}}$.

We implement the following estimation problem using Best Subsets:

$$\min_{w} \ \frac{1}{2}\sum_{j=1}^{J}(\hat{Q}_{.5}(X_j) - X_j'w)^2 \qquad \text{s.t.} \quad ||w||_0 \leq K \tag{10}$$

[5] First, *projected gradient descent* is used to find an approximate solution to the best subsets problem. This procedure is equivalent to using stochastic gradient descent to find regression coefficients, at each step enforcing the sparsity of the solution by zeroing out all but the K largest coefficients. Then, these solutions are used to generate bounds on the MIO formulation of the best subsets problem, which is then solved using an MIO solver. Further detail on this method is contained in Section 6.

Where $Q_t(x) = \inf\{x \ : F(x) \geq t\}$ is the quantile function, and $\hat{Q}_t(x)$ is its empirical analogue.[6] We use the *median* as a Huber-robust estimate of the mean (Huber, 1981).

Note that we no longer explicitly constrain the cardinality of the set $\boldsymbol{S}$: instead, the weight vector w is forced to be K-sparse, which induces K-sparsity of the set $\boldsymbol{S}$. Second, the weights themselves are not meaningful in this problem: we are only interested in the variables that are in fact selected by the model.

2.4.2 Site Selection for the CATE

To solve Problem 3, we implement the following estimation procedure using Best Subsets:

$$\min_{w} \ \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{Q}_t(X_j) - X_j'w)^2 \qquad \text{s.t.} \quad ||w||_0 \leq K \tag{11}$$

Here the goal is to minimize the Weighted Quantile Discrepancy between $f^P(X)$ and $f^{\boldsymbol{S}}$ (Fan et al., 2022). The Weighted Quantile Discrepancy is defined:

$$WQD^2[X, X(\boldsymbol{S}), \boldsymbol{\upsilon}] = \sum_{t=1}^{T} \upsilon_t[Q_t(X) - Q_t(X(\boldsymbol{S}))]^2 \tag{12}$$

Where $\boldsymbol{\upsilon}$ is a weight vector.[7] The idea is to get better minimization of the discrepancy between $f^P(X)$ and $f^{\boldsymbol{S}}(X)$ by quantizing the empirical CDFs, finding the sites that best predict as many quantiles of the population CDF as possible.

In general we have:

$$\begin{aligned}
\int f^P(X) - f^{\boldsymbol{S}}(X)\, dx &= \lim_{||\Delta_{X_t}||\to 0} \sum_{t=1}^{T}[f^P(X) - f^{\boldsymbol{S}}(X)]\Delta_{X_t} \\
&= \sum_{t=1}^{T}[\hat{Q}_t(X) - \hat{Q}_t(X(\boldsymbol{S}))] + \sum_{t=1}^{T}\int_{t-1}^{t}[f^P(X) - f^{\boldsymbol{S}}(X)] - [\hat{Q}_t(X) - \hat{Q}_t(X(\boldsymbol{S}))]\, dx \\
&\approx \sum_{t=1}^{T}[\hat{Q}_t(X) - \hat{Q}_t(X(\boldsymbol{S}))]
\end{aligned}$$

[6] The empirical quantile function is defined $\hat{Q}_t(x) = \inf_x \left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_n \leq x\} > t\right\}$

[7] The idea is that user specified weights allow the researcher to estimate welfare effects: for instance, by choosing a rule like $\upsilon_t = 1$ if $t < \alpha$, the researcher can consider treatment effects for the bottom α^{th} quantiles of the distribution, assuming that covariates are appropriately coded, and some prior knoweldge of the direction of the effects of moderators. In what follows, we take $\upsilon_t = \mathbf{1}_T$.

Or, in words, that the discrepancy is the difference between the sum of the empirical quantile discrepancies plus *quantization error*.

2.5 An Ensemble Method: Subsampling with Weighted Majority Voting

Our goal is to develop a method with improved out-of-sample performance. Ensemble methods, in which the predictions from a base leaner are aggregated to form a final prediction, are a popular tool in machine learning and causal inference that are increasingly used in the social sciences (Zhou, 2012; Montgomery et al., 2012; Samii et al., 2016; Grimmer et al., 2017; Künzel et al., 2019; Grimmer et al., 2021; Athey et al., 2019). The classic best subsets problem was the focus of much early ensemble research, due to the computational intractability of finding exact solutions (Breiman, 1995, 1996). Here, we use ensemble methods as a form of implicit regularization: to prevent overfitting to observed data, and thereby to improve generalization performance. The approach is broadly inspired by the Weighted Majority Voting algorithm of Littlestone and Warmuth (1994).

The general idea is to generate many test-training sample splits, generate site selections for each training set, and evaluate them on our hold out testing set. We then take the observed empirical loss on the testing set, and use this to determine each submodel's contribution to the overall prediction of which sites to use.

The goal is to build robustness to endogenous distribution shifts observed in the data: instead of solving the site selection problem once, we induce distribution shifts by solving the site selection problem for subsets of the data, and then assessing the ability of that prediction to generalize to unseen data. At each iteration, we draw a new test-training split, in which we train the model on $\tau\%$of the data and test it on $1-\tau\%$. We generally set $\tau = .9$.

Formally, let $\hat{S}^{(b)}$ be the prediction of a Best Subsets procedure at an iteration $b \in B$, let $\beta^{(b)}$ be a weight vector, and let H_K be the *hard thresholding operator*, which sets all but the K largest elements of a vector to zero (Donoho and Johnstone, 1994)[8].

Then, our our final site selection for a given set of predictions is the aggregated prediction:

$$\hat{S} = H_K \left[\sum_{b \in B} \beta^{(b)} \hat{S}^{(b)} \right] \tag{13}$$

Let $\ell : \mathcal{X}, \theta \to \mathbb{R}$ be a loss function, $\hat{\ell}(X, \hat{S})$ be the observed empirical loss on our selected sites. Then

[8]The hard-thresholding operator. For a vector $X = \{X_1, \ldots, X_n\}$ take the ordered vector $\{X_{(1)}, \ldots, X_{(n)}\}$. Then, for $K < n$, $H_K(X) = \{X_{(1)}, \ldots, X_{(K)}, 0, \ldots, 0\}$

weights are defined as follows:

$$\beta^{(b)} = 1 - \sqrt{\frac{e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}{1 + e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}}$$

Which is a normalization of the loss: larger losses mean that the contribution of that model to the eventual prediction is smaller.

2.5.1 Loss functions

To motivate our loss functions, we first define two distance metrics:

Definition 3 (1-Wasserstein Distance)

$$W_1(\mathcal{P}, S) = \int_0^1 |F^{-1}_{\mathcal{P}}(\tau) - F^{-1}_{S}(\tau)| \, d\tau$$

Which has a representation as the total absolute distance between the empirical CDFs of the treatment effect in the population and the selected sample. The Wasserstein distance is frequently used in the optimal transport and distributionally robust optimization literatures as a distance metric on the space of distributions (Villani, 2003; Azizian et al., 2023). It has an intuitive interpretation as the smallest amount of 'mass' needed to transform one distribution into another. Using the Wasserstein metric in our loss function is intended to upweight sites that incur the least transportation loss from training to test context. That is, it upweights sites that generalize better from one context to another.

We design a loss function that places weight on both the overall discrepancy between the test set and its selected subset; and on specific quantiles of the test set. This is intended to ensure good performance both overall and with respect to the specific quantile prediction task.

In practice, we choose as our loss function, for a given quantile t:

$$\hat{\ell}(X, X(\hat{S})) = W_1[X, X(\hat{S})] + W_1[\hat{Q}_t(X), \hat{Q}_t(X(\hat{S}))]$$

Where, as above, we set $t = .5$ (i.e., the median) when selecting sites for the PATE, and iterate over a grid of quantiles when selecting sites for the CATE.

Our weights are decreasing functions of these empirical losses. Specifically, we have:

$$\beta_{st} = \begin{cases} 1 - \text{Softmax}\left[\sqrt{\text{logit}^{-1}\hat{\ell}(X, X(\hat{S}))}\right] & s \in \hat{S} \\ 0 & o.w. \end{cases}$$

This imposes a distributional requirement: we require that our chosen sites do well at a weighted majority of quantiles of our observed data.

2.6 Algorithms

Algorithm 1: Best Subsets for the PATE

function BS_PATE (X, $T(X) = \text{median}(X)$, K, B)

Input : Matrix of predictors X, Statistic $T(X)$, Subset size K, Iterations B

Output: Selected site indicators $\hat{S}$

for $b \in B$ **do**

  Split X into $X^{(b)}_{\text{test}}$, $X^{(b)}_{\text{train}}$

  $\hat{S}^{(b)} = \text{Best\_Subsets}(X^{(b)}_{\text{train}}, K)$

  $\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}) \leftarrow W_1[X^{(b)}_{\text{test}}, X^{(b)}_{\text{test}}(\hat{\boldsymbol{S}})]$

  $\beta^{(b)}_s \leftarrow \begin{cases} 1 - \sqrt{\dfrac{e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}{1 + e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}} & s \in \hat{S}^{(b)} \\ 0 & o.w \end{cases}$

end

$\hat{\beta}^K \leftarrow H_K\left[\left\{\sum_{b \in B} \beta^{(b)}_s\right\}_{s=1}^{S}\right]$

$\hat{S} \leftarrow \left(\mathbb{I}\{\hat{\beta}^K_s \neq 0\}\right)_{s=1}^{S}$

Return $\hat{S}$

Algorithm 2: Best Subsets for the CATE

function BBS_CATE (X, T, K, B)

Input : Matrix of site characteristics X, Vector of quantiles T, Subset size K, Iterations B

Output: Selected site indicators $\hat{S}$

for $b \in B$ **do**

  Split X into $X^{(b)}_{\text{test}}$, $X^{(b)}_{\text{train}}$

  for $t \in T$ **do**

    $\hat{S}^{(b)}_t \leftarrow \text{Best\_Subsets}(X^{(b)}_{\text{train}}, \hat{Q}_t(X), K)$

    $\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}_t) \leftarrow W_1[X^{(b)}_{\text{test}}, X^{(b)}_{\text{test}}(\hat{\boldsymbol{S}})]$

    $$\text{loss}^{(b)}_{st} \leftarrow \begin{cases} \sqrt{\dfrac{e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}_t)}}{1 + e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}_t)}}} & s \in \hat{S}^{(b)}_t \\ 0 & o.w. \end{cases}$$

  end

  for $t \in T$ **do**

    $$\beta^{(b)}_{st} \leftarrow \begin{cases} 1 - Softmax(\text{loss}^{(b)}_{st}) & s \in \hat{S}^{(b)}_t \\ 0 & o.w \end{cases}$$

  end

end

$\hat{\beta}^K \leftarrow H_K \left[\left\{ \sum_{b \in B} \sum_{t \in T} \beta^{(b)}_{st} \right\}_{s=1}^{S} \right]$

$\hat{S} \leftarrow \left(\mathbb{I}\{\hat{\beta}^K_s \neq 0\} \right)_{s=1}^{S}$

Return $\hat{S}$

3 Empirical applications

To assess our method, we run simulations based on three studies from political science, economics, and public health (Hill, 2011; Auerbach and Thachil, 2018; Louizos et al., 2017). Our goal is to assess both in-sample performance and out-of-sample performance. That is, how well the selected sites describe the observed covariate data, under the assumption of no unobserved moderators; and how well the sites describe unobserved sites. Out-of-sample performance gives us some sense of how well the observed sites generalize to unseen examples.

Throughout, we compare the performance of our model selection tool to a the Synthetic Purposive Sampling method proposed by Egami and Lee (2024). We also calculate an infeasible oracle estimate, based on unobserved, synthetic treatment effects, generated under the model

3.1 Settlements: Auerbach and Thachil, 2018

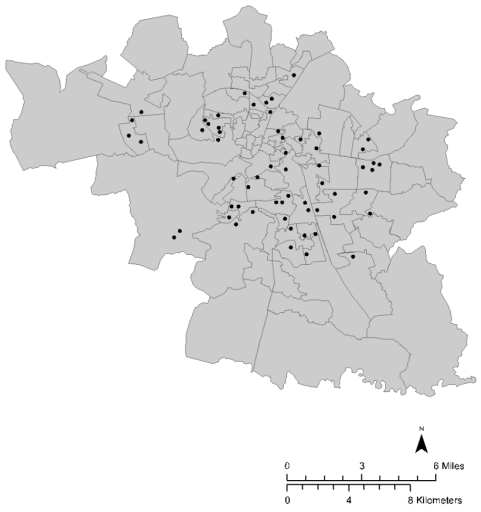

Figure 1: Bhopal

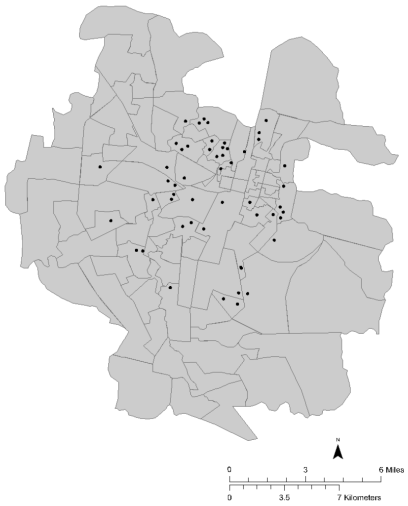

Figure 2: Jaipur

We design a naturalistic simulation study, based on Auerbach and Thachil (2018), which allows us to study both the in-sample and out-of-sample performance of our site selection method. Auerbach and Thachil (2018) conducted a conjoint experiment across a sample of 110 informal urban settlements in Bhopal and Jaipur, India. Their goal was to study resident preferences for political brokers, whose role is to create linkages between political parties and local voters. A main finding from their paper is that residents were

around 13.3% more likely to prefer a broker with a bachelor's degree, which Auerbach and Thachil interpret as a measure of a broker's capability to extract resources from the party.

This setting is a good test of our method for several reasons. First, the authors measured a variety of settlement-level covariates, some of which are predictive of the outcome. This allows us to generate synthetic treatment effects as a function of observed covariates to use as the 'ground truth' in our simulation. Second, because the authors conducted a study in 110 settlements, we have a large and well-defined study population. We can then partition this population into an observed subpopulation, and an unobserved larger population, to assess how well the sites selected in the subpopulation represent sites the unobserved larger population. This creates a naturalistic case study of distribution shift: covariate values will necessarily shift across selections into site. These shifts are not necessarily just due to random sampling, either, as we have no guarantee that all sites are identical based solely on observed characteristics. It is plausible that we observe both sampling uncertainty and distributional uncertainty across both contexts (Rothenhäusler and Bühlmann, 2023).

3.1.1 Simulation approach

Our general approach is as follows:

1. Denote the set of all sites as our population, $\mathcal{P}$.
2. Generate ITEs using individual-level covariate data, and treat these ITEs as the ground truth. This gives us both the PATE and the CATE.

   For $b \in B$:

   3. Randomly sample a **subpopulation** of sites $\mathcal{P}^{(b)} \subset \mathcal{P}$. This is taken to be the population of interest, for which the analyst observes aggregated site-level covariate data.
   4. Use a *site selection method* to select a subset of K sites from the subpopulation $\mathcal{P}^{(b)}$
   5. a) **CATE loss**: Record *the empirical 1-Wasserstein distance* between:
      - *In-sample loss*: The (unobserved) distribution of ITEs in the subpopulation and the distribution of ITEs in the selected sample.
      - *Out-of-sample loss* The (unobserved) distribution of ITEs in the population and the distribution of ITEs in the selected sample.

b) **PATE loss**: Record the empirical 1-Wasserstein distance between:

   · *In-sample loss*: The (unobserved) distribution of SATEs in the subpopulation and the distribution of SATEs in the selected sample.

   · *Out-of-sample loss* The (unobserved) distribution of SATEs in the population and the distribution of SATEs in the selected sample.

c) **Oracle loss**: By brute force search, find the site selection that minimizes each of the above losses with respect to *the unobserved distribution of synthetic treatment effects*. This procedure is infeasible in general because treatment effects are not observed, units outside the population are not observed, *and* brute force search is computationally infeasible for larger sample sizes.

6. Aggregate losses across all replications.

3.1.2 Results

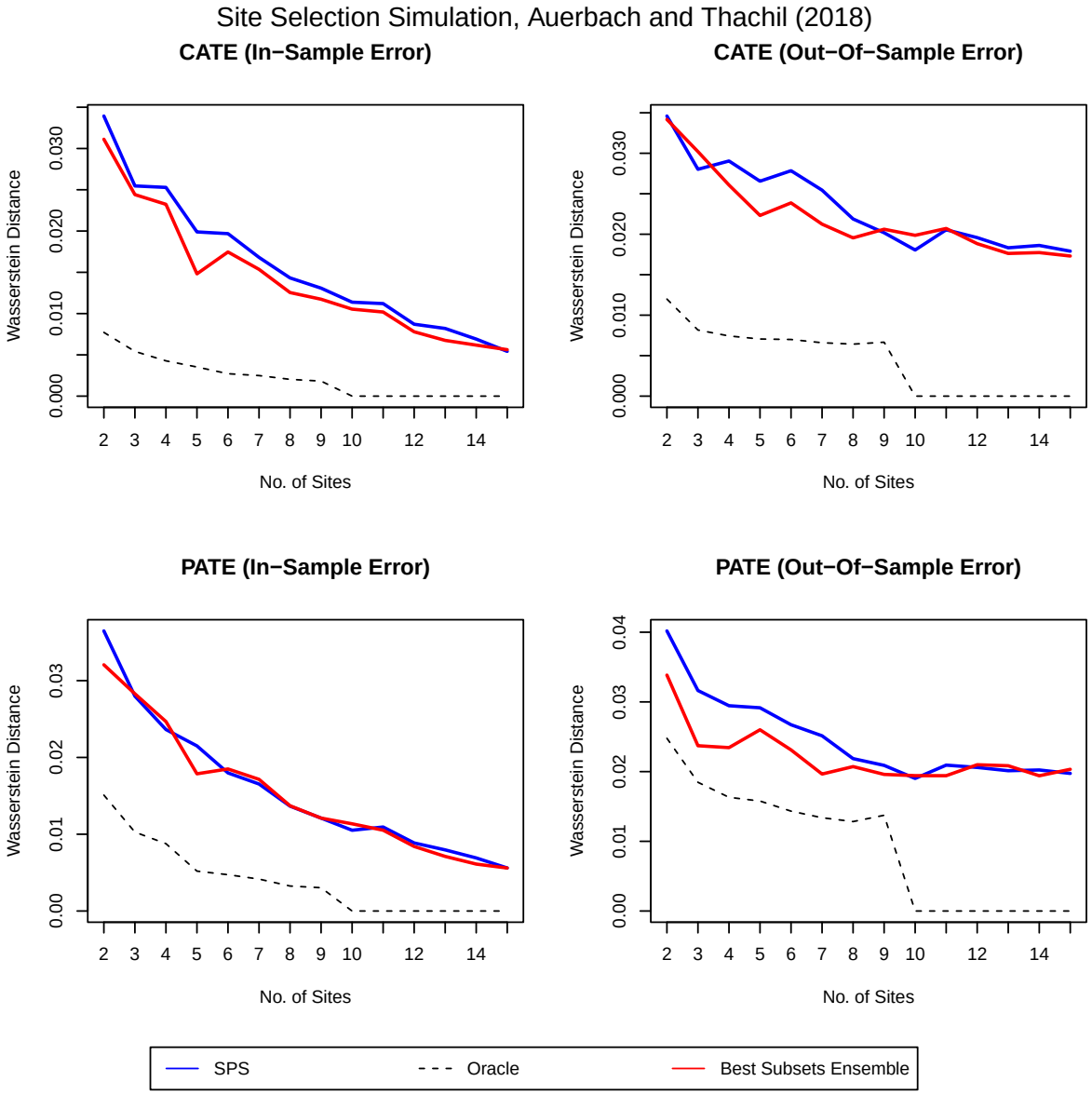

Table 1: Simulation Results, Auerbach and Thachil (2018)

	SPS	Best Subsets Ensemble
PATE (In-sample)	0.01576215	**0.01508495**
PATE (Out-of-Sample)	0.02469087	**0.02217600**
CATE (In-sample)	0.01573571	**0.01396278**
CATE (Out-of-Sample)	0.02332683	**0.02228081**

Our proposed method has better average performance when estimating the PATE and the CATE; and in both in-sample and out-of-sample contexts, compared to Synthetic Purposive Sampling.

3.2 IHDP: Hill, 2011

Our second study uses data from the Infant Health and Development Program (IHDP), which has become a benchmark dataset for assessing the performance of models that estimate heterogeneous treatment effects (Brooks-Gunn et al., 1992; Hill, 2011).

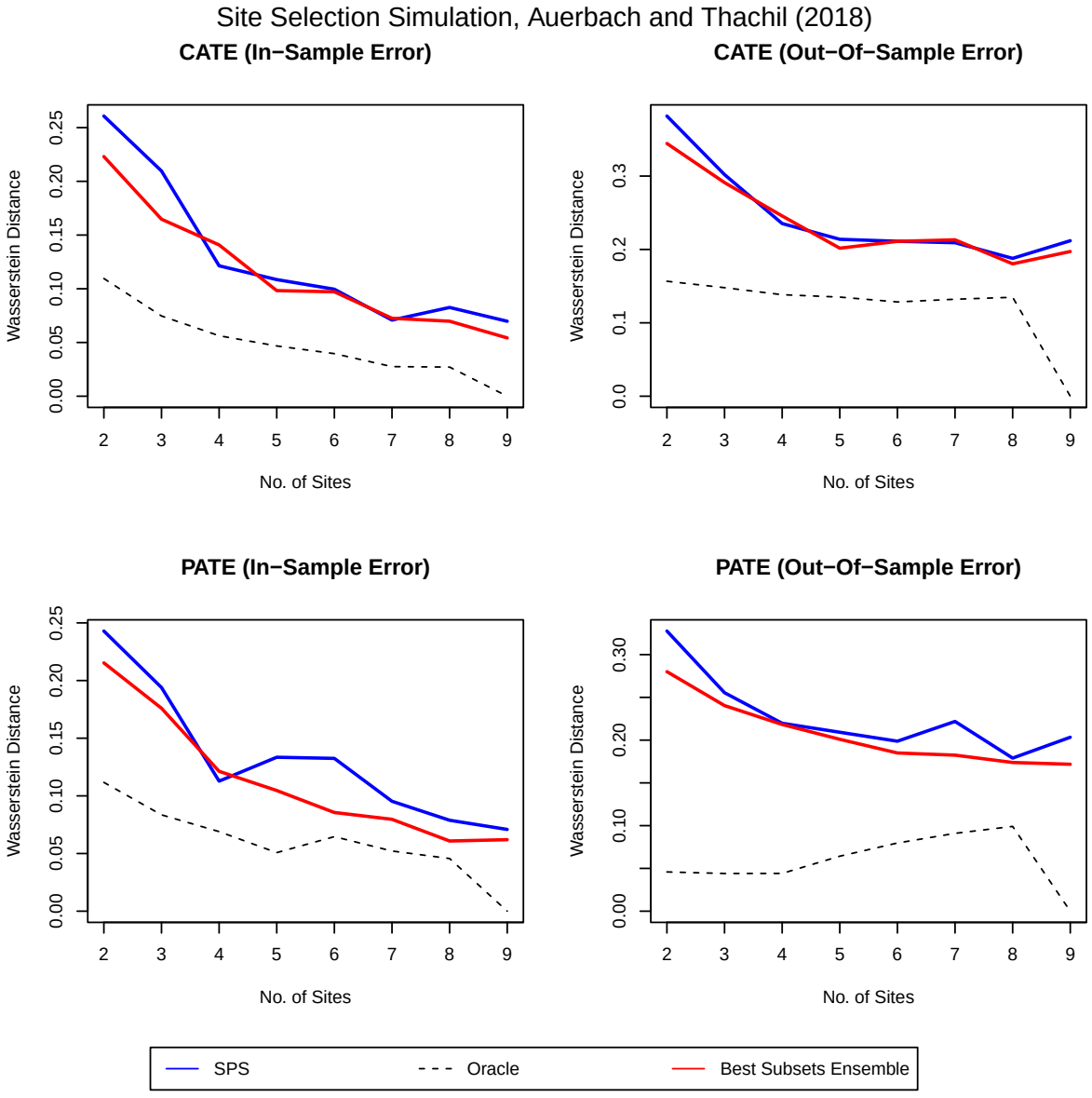

Table 2: Error (Wasserstein Distance), Hill (2011)

	SPS	Best Subsets Ensemble
PATE (In-sample)	0.1325840	**0.1131592**
PATE (Out-of-Sample)	0.2268514	**0.2065952**
CATE (In-sample)	0.1278788	**0.1150756**
CATE (Out-of-Sample)	0.2441207	**0.2355789**

3.3 Twins: Louizos, 2017

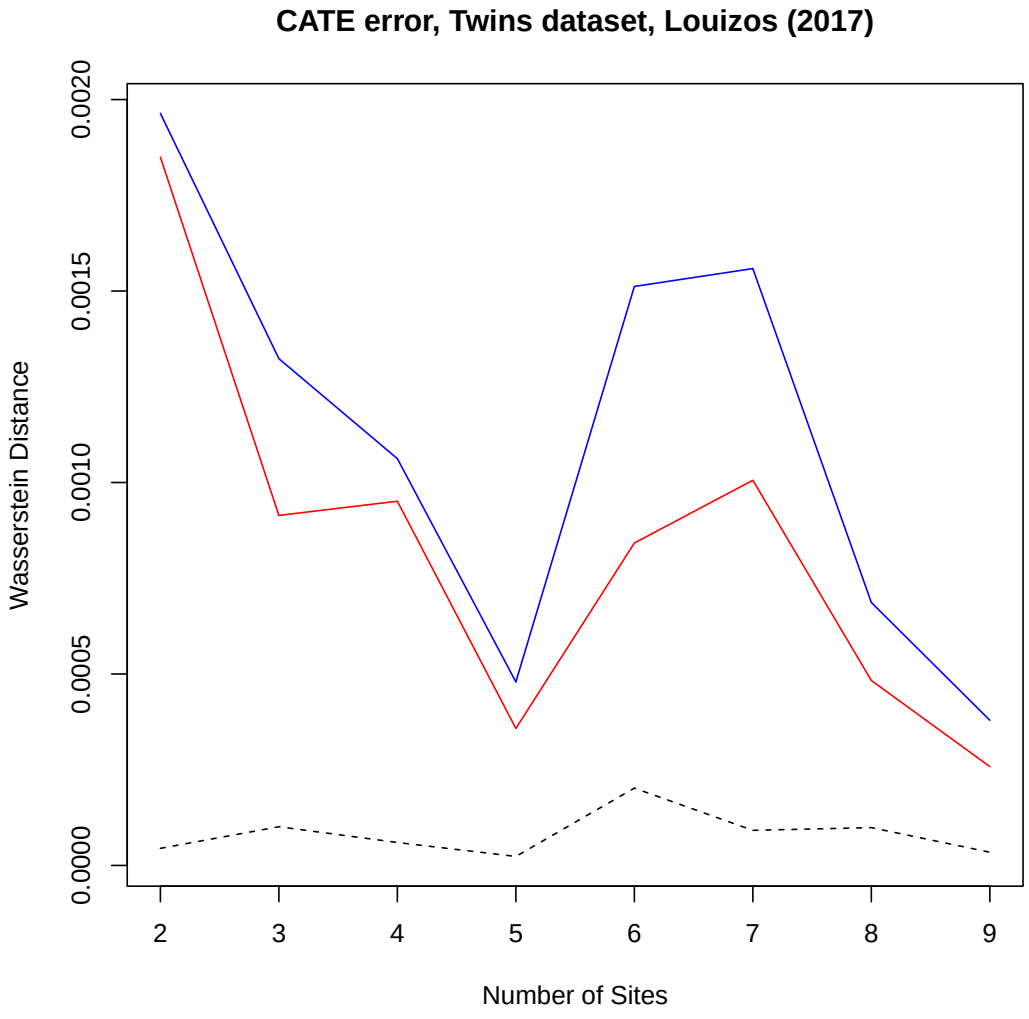

Figure 3: Twins simulation. Blue: SPS, Red: Best Subsets, Black: Oracle

Table 3: Test error (Wasserstein distance) on *Twins* dataset, Louizos (2017)

	SPS	*BestSubsetsEnsemble*
$K = 2$	0.0016831510	**0.0001232097**
$K = 3$	0.0014473130	**0.0006458922**
$K = 4$	0.0014986730	**0.0002082345**
$K = 5$	0.0002831783	**0.0001110586**
$K = 6$	0.0013974690	**0.0003896494**
$K = 7$	0.0023271380	**0.0012880440**
$K = 8$	0.0005380903	**0.0004933354**
$K = 9$	0.0003607147	**0.0001697653**
Mean	0.0011001370	**0.0004286486**

4 Extensions

4.1 Pilot Data

4.2 Welfare Weights

5 Conclusions

We propose a novel method for selecting experimental sites based on observable covariate data using an ensemble learner based on Best Subsets. We show that this method performs well across a range of empirical contexts.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Addanki, R., Arbour, D., Mai, T., Musco, C., and Rao, A. (2022). Sample constrained treatment effect estimation.

Alberto Abadie, A. D. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.

Auerbach, A. and Thachil, T. (2018). How clients select brokers: Competition and choice in india's slums. *American Political Science Review*, 112(4):775–791.

Azizian, W., Iutzeler, F., and Malick, J. (2023). Regularization for wasserstein distributionally robust optimization.

Banaszczyk, W. (1998). Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360.

Bansal, N., Dadush, D., Garg, S., and Lovett, S. (2017). The gram-schmidt walk: A cure for the banaszczyk blues.

Bansal, N., Laddha, A., and Vempala, S. S. (2022). A unified approach to discrepancy minimization.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366.

Beck, J. and Fiala, T. (1981). "integer-making" theorems. *Discrete Applied Mathematics*, 3(1):1–8.

Bertsimas, D., King, A., and Mazumder, R. (2015). Best subset selection via a modern optimization lens.

Bicalho, C., Bouyamourn, A., and Dunning, T. (2022). Conditional balance tests: Increasing sensitivity and specificity with prognostic covariates.

Box, G. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics. Wiley.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Brooks-Gunn, J., ruey Liaw, F., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics*, 120(3):350–359.

Chazelle, B. (2000). *The Discrepancy Method: Randomness and Complexity*. Randomness and Complexity. Cambridge University Press.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science Medicine*, 210:2–21. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.

Ding, P., Feller, A., and Miratrix, L. (2017). Decomposing treatment effect variation.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Duchi, J. and Namkoong, H. (2017). Variance-based regularization with convex objectives.

Duchi, J. and Namkoong, H. (2020). Learning models with uniform performance via distributionally robust optimization.

Egami, N. and Hartman, E. (2023). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, 117(3):1070–1088.

Egami, N. and Lee, D. D. I. (2024). Designing multi-site studies for external validity: Site selection via synthetic purposive sampling. https://naokiegami.com/paper/sps.pdf. [Accessed 15-03-2024].

Fan, Z., Xu, Q., Jiang, C., and Ding, S. X. (2022). Weighted quantile discrepancy-based deep domain adaptation network for intelligent fault diagnosis. *Knowledge-Based Systems*, 240:108149.

Findley, M. G., Jensen, N. M., Malesky, E. J., and Pepinsky, T. B. (2016). Can results-free review reduce publication bias? the results and implications of a pilot study. *Comparative Political Studies*, 49(13):1667–1703.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Gechter, M., Hirano, K., Lee, J., Mahmud, M., Mondal, O., Morduch, J., Ravindran, S., and Shonchoy, A. S. (2024). Selecting experimental sites for external validity.

Grimmer, J., Messing, S., and Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(Volume 24, 2021):395–419.

Harshaw, C., Sävje, F., Spielman, D., and Zhang, P. (2023). Balancing covariates in randomized experiments with the gram-schmidt walk design.

Hartman, E. (2021). *Generalizing Experimental Results*, page 385–410. Cambridge University Press.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579 – 592.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer US.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Levy, A., Ramadas, H., and Rothvoss, T. (2017). Deterministic discrepancy minimization via the multiplicative weight update method.

Li, X. and Ding, P. (2016). General forms of finite population central limit theorems with applications to causal inference.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA. Curran Associates Inc.

Miller, A. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC.

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3):271–291.

Murray, M. K. and Rice, J. W. (1993). *Differential geometry and statistics*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Philadelphia, PA.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234.

Rothenhäusler, D. and Bühlmann, P. (2023). Distributionally robust and generalizable inference.

Samii, C., Paler, L., and Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Number v. 1 in Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin.

Slough, T. and Tyson, S. A. (2023). External validity and meta-analysis. *American Journal of Political Science*, 67(2):440–455.

Sun, L., Ben-Michael, E., and Feller, A. (2023). Using multiple outcomes to improve the synthetic control method.

Thompson, R. (2022). Robust subset selection. *Computational Statistics amp; Data Analysis*, 169:107415.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266.

Tipton, E. and Mamakos, M. (2023). Designing randomized experiments to predict unit-specific treatment effects.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition.

6 Appendix

A.1 Best Subsets and Mixed Integer Optimization

A.2 Proofs of Theorems

A.3 Additional Simulation Results