Optimal Design of Experiments in Panel Data Settings

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the optimal design of experimental studies that have pre-treatment outcome data available. The average treatment effect is estimated as the difference between the weighted average outcomes of the treated and control units. A number of commonly used approaches fit this formulation, including the difference-in-means estimator and a variety of synthetic-control techniques. We propose several different novel estimators and motivate the choice between them depending on the underlying assumptions the researcher is willing to make. Observing the NP-hardness of the problem, we introduce a mixed-integer programming formulation which selects both the treatment and control sets and unit weightings. We prove that these proposed estimators lead to qualitatively different experimental units being selected for treatment. We use simulations based on publicly available data from the US Bureau of Labor Statistics that show improvements in terms of the mean squared error of the estimates when compared to simple and commonly used alternatives such as randomized trials.

1 Introduction

Randomized experiments have long been a staple of applied causal inference. In his seminal paper, Rubin (1974) suggests that "given a choice between the data from a randomized experiment and an equivalent nonrandomized study, one should choose the data from the experiment, especially in the social sciences where much of the variability is often unassigned to particular causes." Using the language of Rubin's *potential outcomes* framework, randomization guarantees that the treatment status is independent of the potential outcomes and that a simple and intuitive estimator that compares the average outcomes of the treatment and control units is an unbiased estimator of the *average treatment effect* (ATE). If both the treatment and control samples are sufficiently large, the hope is that this difference in means estimate is close to the population mean of the treatment effect.

Another crucial property of randomized experimental designs is their robustness to alternative assumptions about the data generating process—a completely randomized experiment does not take into account any features of the observed data. Perhaps not surprisingly, when the researchers are willing to incorporate additional probabilistic assumptions in their design decisions, they can improve on the statistical properties of the average treatment effect estimators (see, for example, Kasy, 2016). These improvements, however, do not come for free and the performance of the estimators may suffer if the incorporated assumptions are violated (Banerjee et al., 2020).

For these reasons randomized experiments are widely used in academic and clinical settings, as well as industrial applications. However, not every practically important question can be easily answered using an experiment on a large sample of experimental units. For instance, the evaluation of major policies targeting large geographic areas has long been of interest in social and political sciences. One of the traditional approaches is to compare the affected unit—such as a metropolitan area or a state—to the average across a sample of carefully picked control units which are deemed to be suitable comparisons (Card, 1990). A more recent alternative, called *synthetic control*, first introduced

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

by Abadie and Gardeazabal (2003) and later developed in Abadie et al. (2010) and Abadie et al. (2015), compares the unit of interest to a weighted average of the units unaffected by the treatment, where the weights are selected in a way that achieves a good fit on pre-treatment outcome variables as well as potentially other observed co-variates.

While originally developed by academics for evaluating the effects of policies, approaches similar to the synthetic control methodology have recently gained popularity in industry as well in cases when applied researchers decide to run experiments targeting larger units often representing geographic areas. This decision may be justified when more granular experiments are either unavailable (for example, television advertising can only be targeted at the media market level) or are unlikely to capture the relevant effects due to interference or equilibrium concerns (Sobel, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008). For instance, a company like Uber may want to evaluate some of the potential treatments at the market level rather than at the driver level if the treatment in question is likely to affect the driver supply. Moreover, launching an experiment targeting even a single unit may be so expensive that the researchers try to minimize the required number of treated units.

Synthetic control and similar approaches may be attractive as estimation procedures in those cases, but they fail to address an equally if not more important aspect such as the optimal choice of experimental units (see, for example, Rubin et al., 2008). We attempt to narrow this gap in the current paper. We consider a panel data setting in which the researcher observes the outcome metric of interest for a number of units in a number of time periods and has to decide: (i) which units to experiment on and (ii) how to estimate the treatment effects after collecting the outcome data in the experimental time periods. The main difference between this setting compared to a typical synthetic control study is that the treated units are not fixed, but rather chosen by the researcher. Proving that the underlying optimization problem is NP-hard, we rule out the possibility of designing a polynomial-time algorithms for the problem under P$\neq$NP. Therefore, we formulate this combined design and analysis problem as a *mixed-integer program* (MIP). Depending on the particular estimands of interest, we propose one of the several formulations and discuss their advantages and drawbacks. We motivate the choice of the optimization objectives and discuss the selection of experimental units each of the objectives leads to. The MIP formulation allows for an easy inclusion of additional constraints as long as those are linear. For instance, it is easy to restrict the overall number of treated units, exclude specific units from treatment, or enforce a budget constraint if there is a varying cost to treat different units.

Using publicly available state-level unemployment data from the US Bureau of Labor Statistics, we compare the proposed methodology to a randomized design that utilizes either the conventional difference-in-means estimator or the synthetic control approach. We estimate the average as well as individual state-level treatment effects in a simulated experiment and find that our approach substantially reduces the *root-mean-square error* (RMSE) of the estimates. We show that our MIP-based design and analysis procedures consistently outperform the more traditional baselines regardless of whether the treatment effects are homogeneous or heterogeneous and whether the number of units selected for treatment is small or large relative to the total number of units in the sample.

The rest of the paper is organized as follows. Section 2 introduces the setting and the proposed estimation approaches. Section 3 presents some of the theoretical unit selection results and the intuition behind them. Section 4 reports the empirical results obtained through simulations. Section 5 discusses some of the practical consideration that should accompany applied work that uses the proposed methodology. Section 6 concludes and outlines the directions of future research.

2 Setting

Let the researcher observe the outcome metric of interest Y for N units during T time periods, such that the observed data can be represented as an $N \times T$ matrix of variables Y_{it}. At time $t = T$ the researcher decides—based on the data observed up until that point—which units should be treated and assigns a binary treatment described by variables $D_i \in \{0, 1\}$, $i = 1, \ldots, N$. The outcomes are then observed for an additional $S - T$ time periods $t = T + 1, \ldots, S$ and the treatment effect estimates are constructed. Each unit $i = 1, \ldots, N$ in each time period $t = T + 1, \ldots, S$ is associated with two potential outcomes $(Y_{it}(0), Y_{it}(1))$, where $Y_{it}(0)$ is realized if $D_i = 0$ and $Y_{it}(1)$ is realized if $D_i = 1$ so that the observed outcome is $Y_{it} = Y_{it}(D_i) = Y_{it}(0)(1 - D_i) + Y_{it}(1)D_i$.

Recall that, given a setting where a single unit $i = N$ has received the treatment in a single time period $t = T + 1 = S$, the synthetic control literature (Abadie et al., 2010, among others) suggests constructing a counter-factual estimate for unit N as a weighted average of the other units' observed outcomes: $\hat{Y}_{N,T+1} = \sum_{i=1}^{N-1} w_i Y_{i,T+1}$. In the previous equation, the w_i's are weights learned from the data observed during the pre-treatment periods $t = 1, 2, \ldots, T$, often by minimizing $\sum_{t=1}^{T}(Y_{Nt} - \sum_{i=1}^{N-1} w_i Y_{it})^2$ under some constraints on the weights. Assuming that the treatment effect, τ, for this unit is additive, it can then be estimated as $\hat{\tau} = Y_{N,T+1} - \hat{Y}_{N,T+1}$.

We now consider a more general setting where K units have received the treatment, with outcomes given by

$$\begin{aligned} Y_{it}(0) &= \mu_{it} + \varepsilon_{it}, \\ Y_{i,T+1} &= Y_{i,T+1}(0) + D_i \tau_i. \end{aligned}$$

with homoscedastic noise ε_{it} that has variance σ^2 and additive treatment effects τ_i. In order to estimate the treatment effect in this more general setting, we can apply a separate synthetic control method to each treated unit i, learning an appropriate set of weights $\{w_j^i\}$ for each treated unit i individually: $\hat{Y}_{i,T+1} = \sum_{j:\, D_j=0} w_j^i Y_{j,T+1}$. The treatment effect for unit i is then estimated as $\hat{\tau}_i = Y_{i,T+1} - \hat{Y}_{i,T+1}$.

Rather than considering each weight-fitting optimization problem separately, we can express the *mean squared error* (MSE) of the resulting estimator of the treatment effect, as:

$$\mathbb{E}[(\hat{\tau}_i - \tau_i)^2] = \left(\mu_{i,T+1} - \sum_{j:\, D_j=0} w_j^i \mu_{j,T+1} \right)^2 + \sigma^2 \left(1 + \sum_{j:\, D_j=0} (w_j^i)^2 \right).$$

The synthetic control literature often operates in settings where the treated units are given. We, however, allow the experimenter to select which units should receive the treatment. The mean squared error formula above leads us to consider the following optimization problem over the weights $\{w_j^i\}_{i,j=1}^N$ and the treatment variables $\{D_i\}_{i=1}^N$ (the objective being the observed sample analog of the right-hand side of the population equation above):

$$\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad \frac{1}{K} \sum_{i=1}^{N} D_i \left[\frac{1}{T} \sum_{t=1}^{T} \left(Y_{it} - \sum_{j=1}^{N} w_j^i (1 - D_j) Y_{jt} \right)^2 + \sigma^2 \sum_{j=1}^{N} \left(w_j^i \right)^2 \right] \quad \text{(per-unit)}$$

under some appropriate constraints on the weights $\{w_j^i\}_{i,j=1}^N$.

In many applications, the researcher may be interested in estimating some weighted average of the unit-level treatment effects on the treated units. Rather than considering each treated unit separately and then computing the weighted average of the estimated individual treatment effects, practitioners may wish to construct a synthetic-control-type estimate for the weighed average of the treated units directly.

Consider a set of treatment assignments $\{D_i\}_{i=1}^N$ and weights $\{w_i\}_{i=1}^N$ on outcomes at time $T + 1$, and assume that we wish to estimate the *weighted average treatment effect on the treated* (wATET), $\tau = \sum_{i:\, D_i=1} w_i \tau_i = \sum_{i=1}^N D_i w_i \tau_i$, as a difference in weighted means: $\hat{\tau} = \sum_{i:\, D_i=1} w_i Y_{i,T+1} - \sum_{i:\, D_i=0} w_i Y_{i,T+1}$. Then, under the same outcomes model as presented above, the mean squared error of the difference-in-(weighted)-means estimator is

$$\mathbb{E}\left[(\hat{\tau} - \tau)^2\right] = \left(\sum_{i:\, D_i=1} w_i \mu_{i,T+1} - \sum_{i:\, D_i=0} w_i \mu_{i,T+1} \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2.$$

As before, our setting allows the experimenter to optimally select which units should receive the treatment as well as which particular weighting scheme should be used. This is especially appropriate when the treatment effects are homogeneous and $\tau_i = \tau$ for all $i = 1, \ldots, N$. In this case, any weighted average of the unit-level treatment effects is equal to τ and the weights can be chosen in a way that minimizes the mean squared error.

The population equation above suggests solving the following optimization problem on the weights $\{w_i\}_{i=1}^N$ and the treatment variables $\{D_i\}_{i=1}^N$ based on the data observed in periods $t = 1, \ldots, T$:

$$\min_{\{D_i, w_i\}_{i=1}^N} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i:\, D_i=1} w_i Y_{it} - \sum_{i:\, D_i=0} w_i Y_{it} \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2. \qquad \text{(two-way global)}$$

So far, we have not considered specific constraints on the weights, $\{w_i\}_{i=1}^N$. If $\sigma^2 > 0$, $w_i = 0$ for all $i = 1, \ldots, N$ is the unique optimal solution to the two-way global problem above if the weights are not constrained in any way. In order to avoid this clearly undesirable solution, we assume that the weights $\{w_i\}_{i=1}^N$ are normalized: $\sum_{i:\, D_i=1} w_i = \sum_{i:\, D_i=0} w_i = 1$. While not strictly required, another reasonable set of constraints motivated by estimating a proper weighted average is requiring all weights to be nonnegative, $w_i \geq 0$ for all $i = 1, \ldots, N$.

Finally, the weights on the treated units may be fixed if a specific weighted average—for example, a simple average with equal weighting—needs to be estimated. This constraint is particularly justified when the treatment effects are heterogeneous and different weighting schemes lead to different estimands. For that reason, we formulate another variation of the global problem:

$$\min_{\{D_i, w_i\}_{i=1}^N} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i:\, D_i=1} Y_{it} - \sum_{i:\, D_i=0} w_i Y_{it} \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2. \qquad \text{(one-way global)}$$

2.1 Mixed-integer formulation

The three optimization problems introduced in the first part of Section 2: the per-unit, the two-way global, and the one-way global problems can all be formulated as mixed-integer programs. We now describe the specific optimization problems we use in the empirical section of the paper.

The per-unit problem is formulated as:

$$\begin{aligned}
\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad & \frac{1}{K} \sum_{i=1}^{N} D_i \left[\frac{1}{T} \sum_{t=1}^{T} \left(Y_{it} - \sum_{j=1}^{N} w_j^i (1 - D_j) Y_{jt} \right)^2 + \lambda \sum_{j=1}^{N} \left(w_j^i \right)^2 \right] \\
\text{s.t.} \quad & w_j^i \geq 0,\ D_i \in \{0, 1\} \ \text{ for } i, j = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \\
& \sum_{i=1}^{N} w_j^i (1 - D_j) = 1 \ \text{ for } i = 1, \ldots, N \text{ such that } D_i = 1.
\end{aligned}$$

The two-way global problem can be formulated as:

$$\begin{aligned}
\min_{\{D_i, w_i\}_{i=1}^N} \quad & \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} w_i D_i Y_{it} - \sum_{i=1}^{N} w_i (1 - D_i) Y_{it} \right)^2 + \lambda \sum_{i=1}^{N} w_i^2 \\
\text{s.t.} \quad & w_i \geq 0,\ D_i \in \{0, 1\} \ \text{ for } i = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \\
& \sum_{i=1}^{N} w_i D_i = 1, \qquad \sum_{i=1}^{N} w_i (1 - D_i) = 1
\end{aligned}$$

and the one-way global problem is simply the two-way global problem with an additional set of constraints: $w_i = w_j$ for all $i, j = 1, \ldots, N$ such that $D_i = D_j = 1$.

All three problems require additional auxiliary variables and constraints in order to have quadratic objectives and only linear constraints (see the Supplementary Materials for the details) so that they

can be conveniently solved using one of the academic or commercial MIP solvers.[1] The term λ that is used in all three objectives is a nonnegative penalty factor. Its selection is discussed in Section 5.

Note that the global versions of the problem can be solved without the constraint on the number of treated units, $\sum_{i=1}^{N} D_i = K$, but the per-unit problem requires it—without the constraint the per-unit problem will tend to select fewer treatment units unless the objective is divided by $\sum_{i=1}^{N} D_i$ which introduces a nonlinearity that we cannot handle. An additional discussion of what that means for fully solving the per-unit problem is provided in Section 5.

3 Design

The three optimization problems introduced in Section 2—the one-way and two-way global and the per-unit problems—tend to select certain treatment units in terms of their location within the distribution of the observed outcome data. In this section we illustrate that behavior using a simple example motivated by the objectives from Section 2. To simplify the analysis, we let $T = 1$ and denote $a_i = Y_{i1}$. We also do not restrict the unit weights to be nonnegative to allow for simpler closed form solutions. Specifically, we consider the following two-way global problem:

$$\begin{aligned} \min_{\{D_i, w_i\}_{i=1}^N} \quad & \left(\sum_{i=1}^{N} w_i D_i a_i - \sum_{i=1}^{N} w_i (1 - D_i) a_i \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2 \\ \text{s.t} \quad & D_i \in \{0, 1\} \text{ for } i, j = 1, \ldots, N, \quad \sum_{i=1}^{N} D_i = K, \\ & \sum_{i=1}^{N} w_i D_i = 1, \ \sum_{i=1}^{N} w_i (1 - D_i) = 1. \end{aligned}$$

This becomes a one-way global problem if we impose an additional set of constraints requiring that $w_i = 1/K$ for all $i = 1, \ldots, N$ such that $D_i = 1$.

A similarly simplified per-unit problem can be written as:

$$\begin{aligned} \min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad & \sum_{i=1}^{N} D_i \left[\left(a_i - \sum_{j=1}^{N} w_j^i (1 - D_j) a_j \right)^2 + \sigma^2 \sum_{j=1}^{N} (w_j^i)^2 \right] \\ \text{s.t} \quad & D_i \in \{0, 1\} \text{ for } i, j = 1, \ldots, N, \quad \sum_{i=1}^{N} D_i = K, \\ & \sum_{j=1}^{N} w_j^i (1 - D_j) = 1 \text{ for } i = 1, \ldots, N. \end{aligned}$$

When the unit weights are optimized (see the Supplementary Materials for the derivations), the optimal values of the objectives can be written in closed-form as functions of the set of treated units, I.

Theorem 1. *Let I denote the set of treatment units and $\bar{I} = \{1, \ldots, N\} \setminus I$ denote the set of control units. Also let $\overline{a}_I = \sum_{i \in I} a_i / |I|$ be the average outcome within the treatment group, $V_I^2 = \sum_{i \in I} (a_i - \overline{a}_I)^2$ a quantity proportional to the sample variance of the outcomes within the treatment group, and the corresponding quantities for set $\bar{I}$ defined similarly. After the unit weights are optimized away, the two-way global objective can be written as:*

$$J_{\textit{2-way}}(I) = \sigma^2 \left(\frac{1}{K} + \frac{1}{N - K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2}{\sigma^2 + V_I^2 + V_{\bar{I}}^2} \right),$$

the one-way global objective can be written as:

$$J_{\textit{1-way}}(I) = \sigma^2 \left(\frac{1}{K} + \frac{1}{N - K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2}{\sigma^2 + V_{\bar{I}}^2} \right),$$

[1] We use SCIP (Gamrath et al., 2020) when generating the empirical results in Sections 3 and 4.

and the per-unit objective can be written as:

$$J_{\textit{per-unit}}(I) = \sigma^2 \left(\frac{1}{K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2 + V_I^2}{\sigma^2 + V_{\bar{I}}^2} \right).$$

It is evident from the objectives that all problems aim to select units in such a way that averages of the observed outcomes for the treatment and control groups are similar—the squared difference between the average outcomes within each group appears in the numerators of all three objectives. The problems differ in terms of how they treat the sample variances of the two groups. The two-way problem attempts to maximize both (taking into account that they are, of course, interdependent)—both quantities appear in the denominator. The one-way problem maximizes the sample variance of the outcomes for the control units only, as the weights for the treated units are fixed—only the term $V_{\bar{I}}^2$ appears in the denominator. The per-unit problem maximizes the sample variance of the control units while minimizing that of the treated units—the term $V_{\bar{I}}^2$ appears in the denominator while the term V_I^2 appears in the numerator. The intuition for the latter is that the per-unit objective tries to simultaneously model each treated unit with a combination of control units, while keeping weight variance small, so it is best to keep treated units as homogeneous as possible.

To illustrate this visually, we solve a two-way global problem and a per-unit problem for a simulated dataset with $N = 25$ units, $T = 2$ pre-treatment periods and the outcomes Y_{it} drawn from a standard normal distribution independently across i and t. This allows 2-d plotting of the units in the space of observed pre-treatment outcomes (Y_{i1}, Y_{i2}). We select either $K = 3$ or $K = 25 - 3 = 22$ treated units and we use $\lambda = \sum_{i=1}^{N} (Y_{i1} - Y_{i2})^2/(4N)$ (which is the average two-period sample variance across all units). Figure 1 shows the units selected by each of the problems.

The results align with the intuition presented using the simple model with $T = 1$ above. Specifically, since the control and the treatment units have symmetric roles in the two-way global objective, the three treatment units selected by the problem when $K = 3$ are the same as the three control units selected when $K = 22$. The control and treatment groups have similar average outcomes and both groups are relatively spread out. The per-unit problem maximizes the spread of the control units which is particularly apparent from the plot corresponding to $K = 22$ while keeping the treatment units as close to each other as possible (see the plot for $K = 3$).

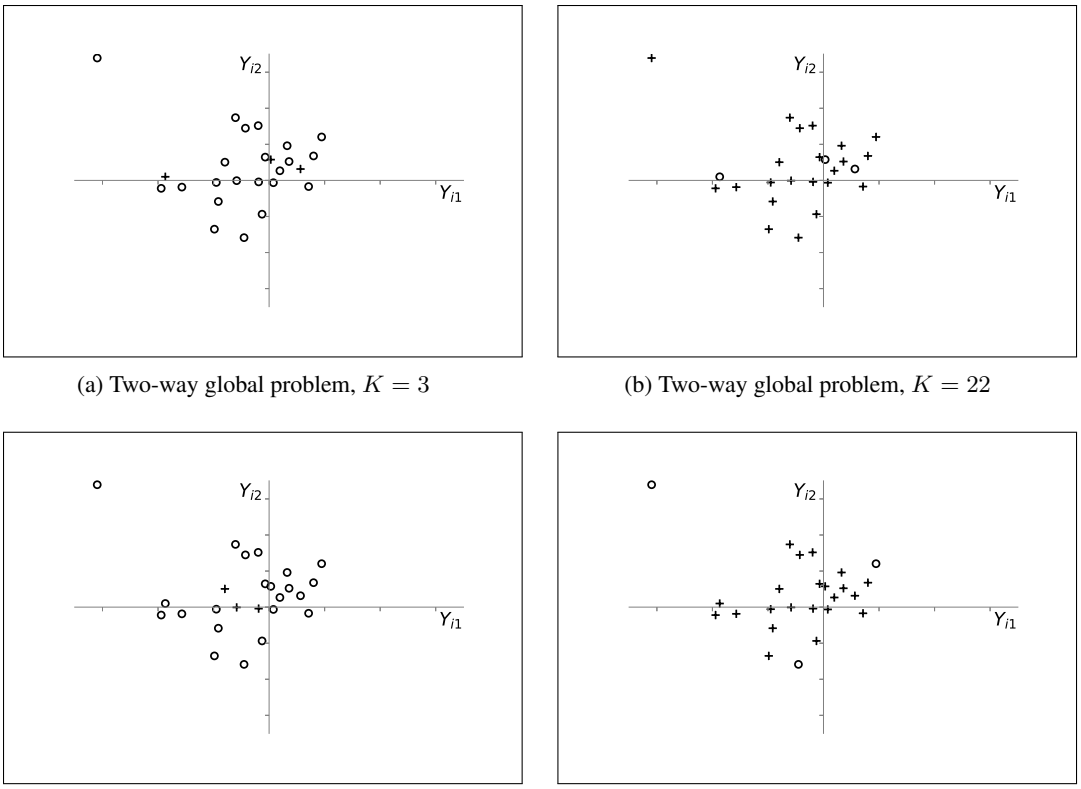
4 Empirical results

To evaluate the performance of the methods proposed in Section 2 we compare several design and estimation procedures: (i) the two-way global problem, (ii) the one-way global problem, (iii) the per-unit problem, (iv) the per-unit problem with randomly chosen treatment units, and (v) the standard randomized experiment which randomly assigns the treatment and estimates the average treatment effect as the difference in means between the two groups. It is important to note that approach (iv) is equivalent to using the synthetic control method[2] for each randomly chosen treatment unit separately and then either averaging the unit-level treatment effect estimates or using the individual estimates directly. Taking this into account, comparing (iii) to (iv) amounts to evaluating the role of optimal design in a synthetic control study, while comparing (iv) to (v) evaluates the synthetic control approach relative to the difference-in-means estimator.

To run a number of simulated experiments, we take publicly available data from the US Bureau of Labor Statistics which contain unemployment rates of 50 states in 40 consecutive months.[3] The dataset is split into a number (63, to be exact) of overlapping 10-by-10 matrices, where each matrix is a block within the original matrix shifted either row-wise or column-wise (or both) with the step size of 5. Each 10-by-10 matrix corresponds to a separate simulation in which K treatment units (equal to 3 in one set of simulations and 7 in another) are chosen based on the data in the first 7 periods—or chosen randomly in cases (iv) and (v)—and the treatment is applied in the last 3 of the 10 periods. We either assign each treated unit the additive treatment effect of 0.05 (homogeneous

[2]The only difference compared to the traditional synthetic control methodology used, for example, in Abadie et al. (2010) is that no additional covariates are used and the weights are obtained using the outcome data alone—the approach similar to the one taken in, for instance, Doudchenko and Imbens (2016).

[3]The data are available from the BLS website, but the specific dataset we use is taken from `https://github.com/synth-inference/synthdid/blob/master/experiments/bdm/data/urate_cps.csv`.



(a) Two-way global problem, $K = 3$

(b) Two-way global problem, $K = 22$

(c) Per-unit problem, $K = 3$

(d) Per-unit problem, $K = 22$

Note: The units are plotted in the space of observed pre-treatment outcome data, (Y_{i1}, Y_{i2}); 'o' denote the selected control units and '+' denote the selected treatment units.

Figure 1: Units selected by the two-way global and per-unit problems.

treatment case) or assume that the first—as determined by the row number—selected treatment unit receives the additive treatment effect of 0, the last unit receives the additive treatment effect of 0.1, and the remaining treatment units receive the treatment effects spread linearly from 0 to 0.1 according to the order determined by their row numbers (heterogeneous treatment case). The average treatment effect on the treated remains 0.05 in both cases.

We estimate the average treatment effect on the treated as well as the unit-level treatment effects. Only the per-unit problem, (iii), and the synthetic control, (iv), allow for nontrivial estimation of heterogeneous treatment effects while the remaining approaches estimate all unit-level effects as being equal to the estimate of the average treatment effect on the treated.

We then compare approaches (i)–(v) in terms of the root-mean-square errors, where the squared differences between the true values of the effects and the respective estimates are computed for each treatment period (and each treatment unit in case of the unit-level effects) and averaged. The square roots of these quantities are the RMSE's in question. Table 1 reports the RMSE's averaged across all simulations.

The specific improvements in terms of the RMSE over the baselines (iv) and (v)—the randomized synthetic control and the randomized difference-in-means—depend on the data and the assigned treatment effects. The main takeaways, however, are more general. The global problems outperform the baselines when estimating the average treatment effect on the treated and the per-unit problem is consistently one of the best when estimating unit-level treatment effects. When the effects are heterogeneous, the one-way global problem provides the most precise estimate of the ATET since the weighting of the unit-level effects coincides with taking the average. For homogeneous treatment effects the difference either disappears, or the two-way approach starts outperforming the one-way approach since any weighting scheme leads to the same value of the ATET.

Table 1: Root-mean-square errors of the average and unit-level treatment effect estimates

Homogeneous treatment				
	$K = 3$		$K = 7$	
	ATET RMSE	Unit-level RMSE	ATET RMSE	Unit-level RMSE
(i) Two-way global	6.9	6.9	7.1	**7.1**
(ii) One-way global	**6.6**	**6.6**	7.3	7.3
(iii) Per-unit	8.2	13.1	**7.1**	14.1
(iv) Synthetic control (random treat.)	8.1	13.2	9.8	16.3
(v) Diff-in-means (random treat.)	9.3	9.3	10.1	10.1
Heterogeneous treatment				
	$K = 3$		$K = 7$	
	ATET RMSE	Unit-level RMSE	ATET RMSE	Unit-level RMSE
(i) Two-way global	7.1	41.6	7.7	34.4
(ii) One-way global	**6.6**	41.5	7.3	34.3
(iii) Per-unit	8.2	**13.1**	**7.1**	**14.1**
(iv) Synthetic control (random treat.)	8.1	13.2	9.8	16.3
(v) Diff-in-means (random treat.)	9.3	42.2	10.1	35.3

Note: The reported RMSE's are multiplied by 10^3 for readability. The values in bold are the lowest in the respective columns and correspond to the methods that perform best.

For the particular simulations we run, the one-way and two-way global objectives provide improvements over the baselines that vary from 12% to 29% in the homogeneous treatment case and from 15% to 29% in the heterogeneous treatment case. The per-unit approach, (iii), performs quite well across the board while being particularly effective when estimating the unit-level effects in the heterogeneous treatment case and providing an improvement of over 13% relative to the synthetic control approach, (iv), when the number of treated units is large. It is not surprising that the per-unit problem provides only a slight improvement over the synthetic control when $K = 3$ because in that case the pool of units that are used for comparison with the treated units is large relative to the overall number of units and the probability that we will not be able to find a good synthetic comparison is relatively low. The situation is different when we only have 3 control units that are used for constructing synthetic outcomes for the 7 treatment units. Section 3 provides an additional discussion of the optimal design when the number of treatment units changes. It is also not surprising that the per-unit approach performs poorly relative to the global problems and even the randomized difference-in-means estimator (but not the synthetic control) when estimating unit-level effects in the homogeneous treatment case. Since all treatment effects are the same, it is more efficient to pool all the data together and estimate the constant effect on the full sample rather than estimating the same quantity for each unit individually. What is important though, is the robustness of the per-unit approach which provides similar performance in the case of either homogeneous or heterogeneous treatment effects while the global problems and the randomized difference-in-means perform very poorly when estimating unit-level effects in the heterogeneous treatment case.

5 Practice

There are a number of practical considerations that need to be addressed when using the proposed design and analysis approaches.

Formulating the mixed-integer programs. Both, the global problem and the per-unit problem can be formulated as mixed-integer programs with quadratic objectives and linear constraints. As discussed in Section 2, the per-unit problem requires an additional constraint that fixes the number of treated units, K. The fact that the per-unit problem has to have the number of treated units fixed means that in general, if the researcher is willing to optimize the number of treated units, the per-unit problem has to be solved the number of times equal to the number of experimental units in the data. In addition to that, the per-unit problem uses more variables that need to be optimized since the weights vary across treatment units allowing separate estimation of every unit-level treatment effect. This implies that the per-unit problem is generally harder to solve and is only tractable for a smaller number of experimental units, N, compared to the global problems.

Choosing the penalty factor. The penalty factor, λ, used by all of the optimization problems can be chosen using cross-validation. Specifically, the pre-treatment time periods can be split into the consecutive training and validation time periods and the value of λ can be chosen by minimizing the RMSE over the validation period in a simulated experiment that is similar to the one we conduct in Section 4. An alternative approach motivated by the setting in Section 2 uses an estimate of the variance of the outcome variable. For example, the approach we take in Sections 3 and 4 computes the sample variances for every unit i across pre-treatment time periods $t = 1, \ldots, T$ and then uses the average of those quantities across all units as the penalty factor, λ.

Quantifying the uncertainty. Most applied settings require evaluating the uncertainty in the obtained estimates of the treatment effects. A careful treatment of this problem is beyond the scope of the current paper, but we suggest using permutation-based inference similar to what is often being done in synthetic control studies (see, for example, Abadie et al., 2010). For instance, the treated units can be temporarily removed from the consideration and the procedure of choice can be applied to subsets of control units as if they were in fact treated. Taking into account the fact that the true treatment effects for those units are equal to zero (since they are not in fact treated), we can generate the distribution of the estimated treatment effects under the null hypothesis of zero treatment effects for all units. The original estimate can then be compared to the quantiles of the constructed distribution.

Computational complexity. Solving the global and the per-unit problems in their mixed-integer formulations becomes computationally burdensome as the total number of units increases especially if the exact optimal solutions are required. In our simulations we were able to solve problems for $N = 50$ units—which is a meaningful threshold corresponding to the number of states, a typical experimental unit in synthetic-control-type studies—on a single machine within hours. However, we prove that the underlying optimization problem is NP-hard (by providing a reduction to the partitioning problem; for the exact proof see the Supplementary Materials), and therefore exact solutions to substantially larger problems are unlikely.

6 Conclusion

In this paper we evaluate the role of optimal experimental design in panel data settings where traditionally the average treatment effect on the treated would be estimated using randomized design and the difference-in-means estimator. We propose several different design and analysis procedures that can be solved as mixed-integer programs. Our empirical evaluations show that these procedures lead to a substantial improvement in terms of the root-mean-square error relative to the randomized difference-in-means as well as the randomized synthetic control approaches.

We discuss how the underlying assumptions about the nature of the treatment effects, the estimands of interest, and the computational considerations should inform the decision to use a particular approach. We also discuss the limitations of the current work.

The are several potential directions for future research that are motivated by the practical considerations raised in Section 5. Perhaps most importantly, in subsequent papers we are planning to explore approximation algorithms that would allow scaling up the applications of proposed methods from relatively small settings with several dozens of experimental units to larger problems.

References

Abadie, A., A. Diamond, and J. Hainmueller (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.

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

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review 93*(1), 113–132.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review 43*(2), 245–257.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Gamrath, G., D. Anderson, K. Bestuzheva, W.-K. Chen, L. Eifler, M. Gasse, P. Gemander, A. Gleixner, L. Gottwald, K. Halbig, G. Hendel, C. Hojny, T. Koch, P. Le Bodic, S. J. Maher, F. Matter, M. Miltenberger, E. Mühmer, B. Müller, M. E. Pfetsch, F. Schlösser, F. Serrano, Y. Shinano, C. Tawfik, S. Vigerske, F. Wegscheider, D. Weninger, and J. Witzig (2020, March). The SCIP Optimization Suite 7.0. Technical report, Optimization Online.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

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

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. et al. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics 2*(3), 808–840.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 2, 3, and 4.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 and Supplementary Materials.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the Supplementary Materials.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We intend to release the code when the methodology proposed in the paper is fully developed for external use and appropriately tested.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sections 4 and 5.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]