Designing Synthetic Control Experiments with Forward Selection

Lawrence R. De Geest
Zoox (formerly Lyft)
San Francisco, CA, USA
ldegeest@zoox.com

Amber Wang
Lyft
San Francisco, CA, USA
awang@lyft.com

ABSTRACT

Synthetic control experiments allow researchers to cleanly estimate both short-term and long-term causal effects of marketplace interventions subject to network effects. However, designing such experiments by selecting marketplaces to treat is challenging because of the size of the feasible design space. We present a simple approximate algorithm based on forward selection that can efficiently pick treated units and estimate treatment effects using any synthetic control method of choice. We explore the behavior and possible modifications of the forward selection algorithm using simulated data assuming a factor model structure.

1 INTRODUCTION

Synthetic control experiments are popular designs for cleanly testing the causal effects of interventions in marketplaces that exhibit interference or "network effects" between users ([2, 7, 9]). Consider a canonical example from ride hailing markets, in which interference emerges from riders competing over a scarce stock of drivers. If we run a price coupon experiment and randomize riders in San Francisco into treatment and control, the treated will buy more rides, but in doing so, interfere with the control by reducing their available stock and worsening service levels (e.g., higher wait times), leading to fewer control rides than had the experiment never run. Because we violate SUTVA and contaminate our counterfactual, this randomized controlled trial or "user-split" will overstate the treatment effect ([11]), and this bias will scale with the elasticity of the supply curve ([12]). When interference bites, one solution is to run a switchback experiment or "time-split", in which we treat and control the entire marketplace at regular intervals. However, this design may also suffer from bias due to carryover effects, and by nature can only provide instantaneous, short-term estimates ([5, 19]).

By contrast, a synthetic-control experiment or "region-split" can capture the externalities created by either or both sides of the marketplace, and allow us to cleanly estimate how treatment effects play out in both the short and the long term. The premise is simple enough. Instead of splitting San Francisco in half, we treat all riders in San Francisco with the coupon and compare the treated market to other marketplaces using a synthetic control estimator.

Motivated by policy evaluation, the academic literature on synthetic controls has largely focused on analyzing region-split experiments, taking the intervention as given. However, organizations in industry are often able to choose regions to treat in the first place. The following are some challenges that can be greatly mitigated by carefully selecting the treated regions:

(1) **Bias**: The fundamental assumption for region-split experiments to recover causal effects is that the control will serve as a good counterfactual for the treated in the near future if the control and the treated exhibit parallel trends in the past. Thus, carefully choosing two sets of regions that historically behave "similarly" is crucial to causal identification.
(2) **Power**: By randomizing over a coarser unit – marketplaces instead of individuals or units of time – we have fewer observations, making it harder to distinguish true effects from noise. If we can select treated regions that ensure the fitted synthetic control is credible, we can be more confident that a post-treatment difference between the treated and control is not due to noise.
(3) **Practical considerations**: business and policy considerations, such as the risk of exposing certain regions to treatment, or treating too many regions, may constrain the feasible set of experimental units, and thus jeapordize external validity of feasible designs.

The literature on experiment design with synthetic controls is nascent, focusing primarily on design objectives (e.g. [2]) cast as NP-hard mixed integer problems to be approached using a solver (e.g. [9]), which can be slow as the number of units increases. In this paper, we explore the use of a simple and fast forward selection algorithm to design synthetic control experiments. This algorithm is similar to the Forward Difference-in-Differences (FDID) estimator by [15] where forward selection is deployed to select control units in the construction of a synthetic control. In our research on experimental design, in which we need to select units to treat as well as control, we flipped the estimator's forward selection on its head, greedily searching for units to treat and constructing synthetic controls in each iteration. We present some high-level results on real-world marketplace data, and then explore our algorithm in more detail in a simulation in which marketplaces are governed by a clustered factor model.

This paper is a preliminary exploration of the forward selection algorithm as well as of the bigger topic of searching for optimal treated units in a region-split experiment. We believe that this area is of both theoretical and applied importance and ripe for contributions.

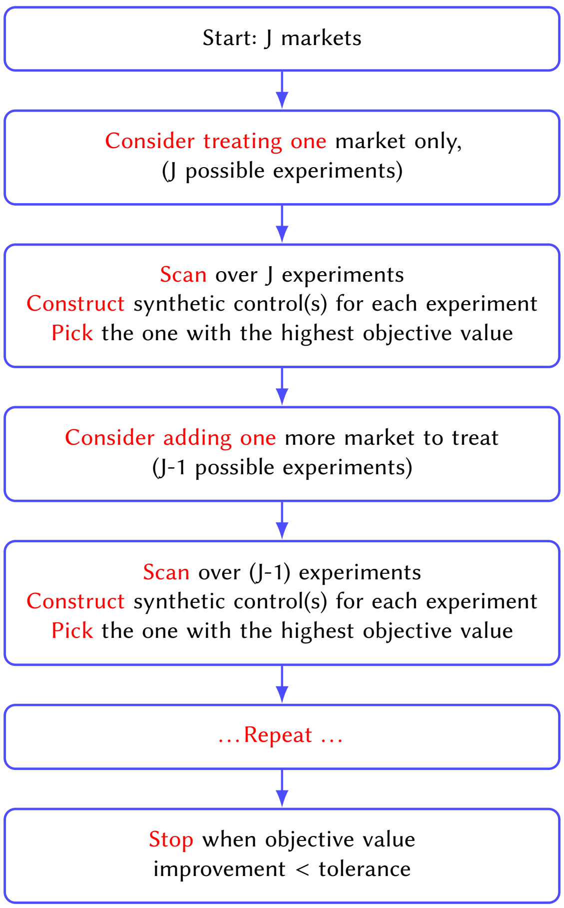
We first present the general forward selection algorithm and give an example of how it can be used with the Forward Difference-in-Differences Estimator. Then we discuss the effect of the algorithm in the context of a factor model with clusters. Finally, we discuss practical considerations for designing synthetic control experiments.

2 DESIGNING WITH FORWARD SELECTION

Previously at Lyft we designed region-split experiments using gradient-free, stochastic optimization techniques like simulated annealing and genetic algorithms ([6]). Both techniques offered improvements over random selection of treated and control units, but they required a significant amount of fine-tuning and are in general slow, especially when one has a relatively large set of regions.

Our proposed greedy search algorithm – forward selection – is fast because we do not aim for an exact solution to the NP-Hard optimization problem the way other papers do. In essence, we start with

J marketplaces or regions, consider each unit $j \in J$ as the treated, pretending we were only to treat one unit (e.g, region j is the treated, and all remaining regions $-j$ are potential donors), construct a synthetic control, check the value of the objective function (e.g. pre-treatment fit), and then pick the best one for the first treated unit. We then iteratively add more treated units to the experiment, one at a time, until the improvement to the objective function value from treating any additional unit becomes negligible (i.e. falls below some pre-determined tolerance). Our approach is visualized by the flow chart below:



One must choose some objective function and synthetic control estimator to design the experiment. Below, we provide a specific example of combining the forward selection algorithm with the objective function optimizing pre-treatment fit and the FDID estimator. There are three benefits to FDID: (1) It runs fast: because a synthetic control is built for each potential set of treated units evaluated during the process, by using forward selection to select donors for the synthetic control, it improves the speed of the forward selection of the treated. (2) It reduces risk of overfitting: the synthetic control is constructed as the simple average across donors plus an intercept. The simple average serves as a coarse regularization of the weights. (3) It provides a closed-form variance estimator: this property only requires weak stationarity of pre-treatment differences, an assumption we find holds well in practice. There are four simple steps to constructing an FDID estimate tabulated below:

Step	Formula	Description
(a)	$y_{S,t} = \frac{1}{N_S}\sum_{s=1}^{S} y_{s,t}$	Construct synthetic control as a simple average over donors.
(b)	$y_{T,t} = \alpha + y_{S,t} + \varepsilon_t$	Model (OLS) the treated as the synthetic control plus an intercept.
(c)	$R^2_S = 1 - \frac{\sum_{t=1}^{T_1}\left(y_{D=1,t} - y_{S,t} - \hat{\alpha}\right)^2}{\sum_{t=1}^{T_1}\left(y_{D=1,t} - \bar{y}_{tr}\right)^2}$	Select donors by maximizing pre-treatment R^2.
(d)	$\hat{\tau} = \frac{1}{T_1}\sum_{t=1}^{T_1}\left(y_{D=1,t} - y_{S,t}\right)$	Estimate the treatment effect.

Table 1: Experiment analysis with FDID (Li, 2024).

We can then formalize our design process in the algorithm below:

Algorithm 1 Greedy Forward Selection for Treated Units

Require: Pre-treatment data $X_{\text{pre}} \in \mathbb{R}^{T\times N}$, outcome $y_{\text{pre}} \in \mathbb{R}^{T}$, tolerance τ
Ensure: Selected treated units $\mathcal{T}^*$
Initialize: Candidate set $C \leftarrow \{1,\ldots,N\}$, selected treated units $\mathcal{T} \leftarrow \emptyset$, previous best $R^2_{\text{best}} \leftarrow 0$
for all $i \in C$ **do**
  $R^2_i \leftarrow 1 - \frac{\sum_t (y_{\text{pre},t} - X_{\text{pre},t,i})^2}{\sum_t (y_{\text{pre},t} - \bar{y}_{\text{pre}})^2}$
end for
$t_1 \leftarrow \text{argmax} R^2_i$
$\mathcal{T} \leftarrow \{t_1\}, \quad C \leftarrow C \setminus \{t_1\}$
Store history: $\mathcal{H} \leftarrow [\mathcal{T}], \quad \mathcal{R} \leftarrow [R^2_{t_1}], \quad R^2_{\text{best}} \leftarrow R^2_{t_1}$
for $k = 2$ to N **do**
  for all $i \in C$ **do**
    $y_{\text{pred},t} \leftarrow \frac{1}{|\mathcal{T}\cup\{i\}|}\sum_{j\in\mathcal{T}\cup\{i\}} X_{\text{pre},t,j}$
    $R^2_i \leftarrow 1 - \frac{\sum_t (y_{\text{pre},t} - y_{\text{pred},t})^2}{\sum_t (y_{\text{pre},t} - \bar{y}_{\text{pre}})^2}$
  end for
  $t_k \leftarrow \text{argmax} R^2_i, \quad R^2_{\text{new}} \leftarrow R^2_{t_k}$
  if $R^2_{\text{new}} - R^2_{\text{best}} < \tau$ **then**
    Stop: Tolerance condition met, exit loop
    Break
  end if
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{t_k\}, \quad C \leftarrow C \setminus \{t_k\}$
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathcal{T}\}, \quad \mathcal{R} \leftarrow \mathcal{R} \cup \{R^2_{t_k}\}$
  $R^2_{\text{best}} \leftarrow R^2_{\text{new}}$
end for
Find the iteration with the highest R^2: $\mathcal{T}^* \leftarrow \mathcal{H}[\text{argmax}\mathcal{R}]$
return $\mathcal{T}^*$

Although we used the FDID estimator in our example, the forward selection algorithm can be deployed with different synthetic control estimators and objective functions, as long as one uses the same estimator in the design of the experiment and in the analysis of the experiment.

2.1 Real-world example

Figure 1 shows some results from our design process using Lyft data. After selecting our treated units, we run one AA test using the synthetic control constructed by the FDID estimator to visualize the fit (left panel of figure). With these treated and control units in hand – a "design" – we then simulate many AA tests using FDID on the same treated and control units while permuting the treatment start dates to check the false-positive rate (FPR) (right panel). The forward selection design outperforms a random design in pre-treatment fit.

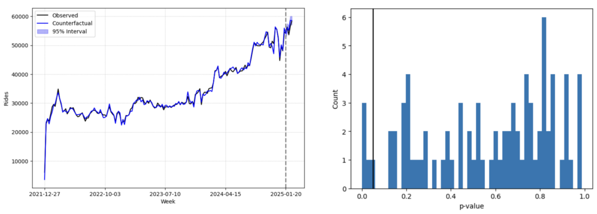


Figure 1: Fitted optimal design and distribution of p-values from placebo tests (false-positive rate).

3 SIMULATION

We explore our experiment design technique with a simple simulation. Consider N regions or marketplaces grouped into G clusters. The outcome for region i in time t is generated by a factor model

$$Y_{it} = \mu + \alpha_{g(i)} + \Lambda_i F_t + \epsilon_{it}$$

where μ is a global intercept, $\alpha_{g(i)}$ is a cluster-random intercept that captures systematic mean differences (e.g., average ride volume), $F_t \in \mathbb{R}^K$ are common factors that represent time-varying details like market conditions and evolve according to a stationary AR(1) process $F_t = \phi F_{t-1} + \eta_t$, and $\Lambda_i \in \mathbb{R}^K$ are the factor loadings for region i.

To introduce some realistic structure to our panel, we allow each cluster $g \in G$ to have a center in the factor-loading space

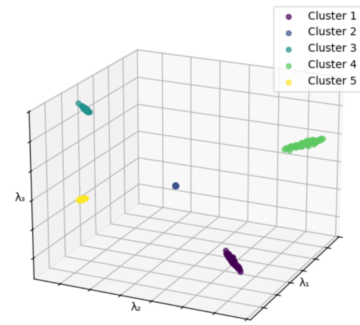
$$c_{g(i)} \sim N(0, \sigma^2_{\text{between}} I_K)$$

and each member of g draws its own loading around that center:

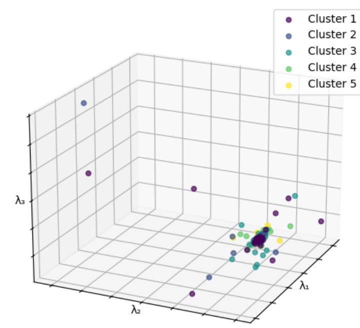
$$\Lambda_i = c_{g(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2_{\text{within}} I_K).$$

thus introducing block-structure in the covariance matrix of Y_{it} and capturing systematic patterns (e.g., structural differences in how regions respond similarly to shocks).

We first explore experiment designs under various degrees of between- versus within-variance. Intuitively, a large between-variance (relative to within-variance) will spread clusters apart in factor space, whereas a small between-variance (relative to within-variance) will move clusters closer together, as illustrated in the figure below.



(a) Large between-variance (relative to within-variance).



(b) Small between-variance (relative to within-variance).

Figure 2: A comparison between large and small between-variance in factor space for $N = 500$ regions.

Figure 3 shows results from simulations that ramp up the between-variance relative to the within-variance. We find that the number of regions forward selection picks to form an experiment (including the treated units and the control units) increases as the between-variance decreases. When between-variance is sufficiently large, the algorithm will only select one region to treat, and all control units will come from the same cluster as the selected treated unit.

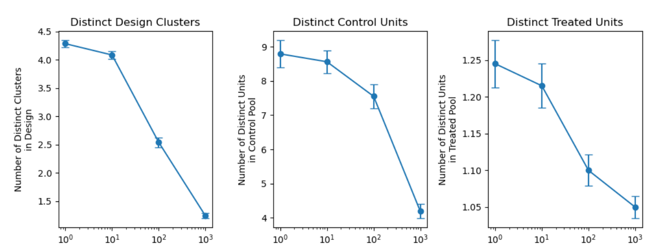


Figure 3: Optimal designs with increasing between-variance relative to within-variance.

One straightforward benefit to forward selection is an increase in power relative to random designs (that either use all available units or that use the same total number of regions resulting from forward selection), as shown in Figure 4.

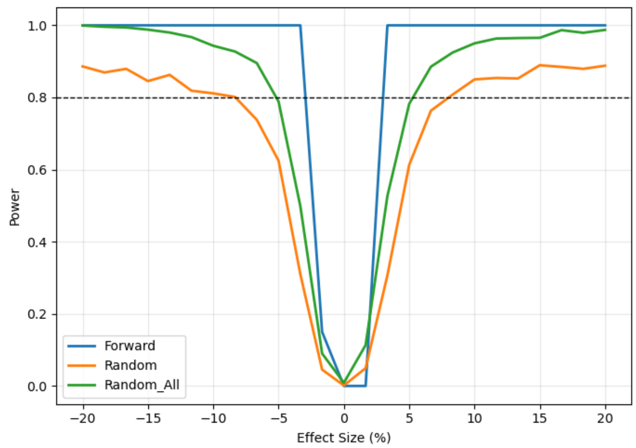


Figure 4: Example power curves for forward selection and random designs under high between-variance (30 marketplaces organized into five equal-sized clusters).

Another benefit to forward selection is that it can produce parsimonious and powerful designs that consume fewer units than random designs. Such an experiment has the upside of incurring very little cost or risk. However, it also has the obvious downside of lacking external validity, especially considering that the small number of selected treated units (e.g. one) is a symptom of large structural differences across regions as captured by the large between-variance.

In this case, one can consider running forward selection sequentially: do one round of forward selection, exclude the selected treated and control regions, and do another round of forward selection using the remaining regions. This process can be repeated until the value of the objective function falls below a certain tolerance. In our simulation, when the between-variance is very large, this sequential forward selection will produce experiments that a cluster-by-cluster design produces, without explicitly knowing the underlying cluster structure. In general, the possibility of doing forward selection sequentially allows one to move along the trade-off frontier between a parsimonious experiment and external validity.

We then explore idiosyncratic noise ϵ_{it}. Specifically, we inject a true treatment effect of 10% and increase idiosyncratic noise, then run 100 simulations for each level holding everything else constant. As Table 2 shows, the mean of the treated size across simulations increases with the idiosyncratic noise and the treatment effect τ estimated by the separate approach becomes less accurate. Intuitively, the fit between the treated and the synthetic control becomes worse as underlying structure is lost in noise. Pooling treated units together to construct one synthetic control performs much better than separately building a synthetic control for each treated unit. The forward selection will also select more treated units if one allows for the pooling approach. This is because averaging across the treated units helps to smooth out the idiosyncratic noise and regions can then be matched based on their true underlying structures.

Table 2: Comparison of pooling and separate treatment effect estimates τ under varying idiosyncratic noise $\sigma^2_{\text{between}}$ with average treated size $\bar{J}_{\text{treated}}$ and design size $\bar{D}$ (number of treated units plus the number of control units).

σ_2	$\bar{J}_{\text{treated}}$	$\bar{D}$	τ_{pooling}	τ_{separate}	$Var(\tau_{\text{separate}})$
5	2.02	12.03	0.094	0.081	3.377
15	2.61	12.02	0.099	0.087	3.513
30	3.68	13.59	0.095	0.077	5.561
50	4.20	16.59	0.135	0.032	15.606

4 TECHNICAL NOTES

4.1 Dealing with multiple treated units

Often there is interest in treating multiple marketplaces or regions, and forward selection often returns such designs. There are in general two approaches to deal with synthetic control estimation in the case of multiple treated units:

(1) The "separate" approach, where the average treated effects of the treated (ATT) is estimated by building one synthetic control for each of the treated units and then taking average across the separately estimated treatment effects. Hence, one experiment with N treated units would require N synthetic controls.
(2) The "pooling" approach, where only one synthetic control is built for the average (or some weighted average) of the treated units (e.g. [14], [16]). [3] show that the error of the Average Treatment Effect on the Treated (ATT) estimate decomposes into error stemming from the pooled fit and from unit-specific fits.

The pooling approach has the advantage of being able to create a better fit because it smooths out idiosyncratic noise. However, [13] suggest that doing so introduces an interpolation bias, which is non-zero if and only if post-intervention outcome is not a linear function of pre-intervention outcomes. [10] propose a "Linearity in Pre-Treatment Outcomes (LPO)" assumption and suggest that it is sufficient to consider the simple average of the treated units under this assumption. They also show that two-way fixed effect (DiD) models, auto-regressive models with local trends, and factor models (a theoretical data-generation process that justifies synthetic control models) all satisfy their version of LPO, which is a weaker version of the linearity requirement in [13].

In practice, whether it is advisable to take the pooling approach depends on the assumption the practitioner is willing to make and the specific goal of the estimation exercise. Alternatively, [1] and [4] suggest penalizing the unit-specific imbalance for each treated unit. This is an adjustment of the objective function and the forward selection algorithm can be used in combination with this "partially pooled SCM" as well.

4.2 Optimality of forward selection

The forward selection algorithm is a greedy search, and thus does not guarantee the "optimal" design is indeed globally optimal. It is possible that a particular treated unit selected early in the process is in fact a critical unit to construct a good synthetic control in a later step. We have two thoughts on this matter.

In the context of region-split experiments, it is usually easy to check whether the performance of the forward selection solution is "good enough". For example, if one cares about power, one can verify whether the minimal detectable effect implied by the forward selection solution is acceptable.

Second, it is easy to extend forward selection by adding a "swapping" step in the end. In this step, the researcher "releases" one or more early selected treated units back into the donor pool and proceeds with forward selection again with the remaining treated units. Many other extensions and modifications are also plausible, and we regard this topic as a promising avenue for future research.

5 PRACTICAL NOTES

First, in experiment design practice, the designer may wish to include or exclude some particular regions due to other considerations. The forward selection algorithm can easily adapt to this demand. One can start with a set of pre-selected treated units and use the forward selection to select additional regions to treat. Some regions can also be excluded from the list of regions to be scanned by the forward selection algorithm.

Next, we have experimented with designs on different time frequencies and have found that weekly data produced the most stable designs, since it absorbs day-to-day variation and provides more data points along the time dimension within a reasonable length of history compared to monthly data.

Lastly, additional covariates can be incorporated by first residualizing the outcome and then running the algorithm and the estimator on the residualized outcomes ([8]).

6 CONCLUSION

Synthetic control estimators are well-established as tools to analyze experimental or quasi-experimental data in the presence of externalities or network effects. However, the design of synthetic control experiments (or "region-split") is a nascent and exciting topic of research. Existing literature mostly discusses different design objectives, leaving the issue of how to solve the implied optimization problem under-explored. Here we have presented a simple and fast greedy search algorithm, namely the forward selection algorithm, to choose units to treat according to a given objective function. This algorithm offers a tractable and interpretable way to operationalize synthetic control design in practice. We have also presented a way to simulate data to explore how the algorithm behaves as the true data generating process changes. We sincerely hope to see more research in this area.

One direction that we have not explored is the case of multiple metrics of interest correlated to varying degrees. There has been some attention given on synthetic control estimators that exploit multiple outcomes (e.g., [18]). Efficiently designing synthetic-control experiments with multiple metrics in mind is an interesting topic for future research.

Another promising avenue for future work is the design of staggered roll-out synthetic control experiments ([4, 17]). In such a design, a researcher would aim to select not only *which* units to treat (and which to control) but *when* to treat them in a sequence. Such designs give organizations greater chances at running informative experiments that permit a sequential decision-making process to continue or cease rolling out the intervention depending on intermediate results.

REFERENCES

[1] Alberto Abadie and Jérémy L'hour. 2021. A penalized synthetic control estimator for disaggregated data. *J. Amer. Statist. Assoc.* 116, 536 (2021), 1817–1834.

[2] Alberto Abadie and Jinglong Zhao. 2021. Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196* (2021).

[3] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. 2021. The augmented synthetic control method. *J. Amer. Statist. Assoc.* 116, 536 (2021), 1789–1803.

[4] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. 2022. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84, 2 (2022), 351–381.

[5] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. 2023. Design and analysis of switchback experiments. *Management Science* 69, 7 (2023), 3759–3777.

[6] Yan Chen, Martin Loncaric, Bobak Moallemi, and Sean J. Taylor. 2020. *Synthetic Control Estimators in Practice.* Technical Report. Lyft.

[7] Nick Doudchenko, David Gilinson, Sean Taylor, and Nils Wernerfelt. 2019. *Designing experiments with synthetic controls.* Technical Report. Technical report, Working paper.

[8] Nikolay Doudchenko and Guido W Imbens. 2016. *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.* Technical Report. National Bureau of Economic Research.

[9] Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, et al. 2021. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems* 34 (2021), 8691–8701.

[10] Chad Hazlett and Yiqing Xu. 2018. Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *Available at SSRN 3214231* (2018).

[11] David Holtz, Felipe Lobel, Ruben Lobel, Inessa Liskovich, and Sinan Aral. 2025. Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on airbnb. *Management Science* 71, 1 (2025), 390–406.

[12] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. 2022. Experimental design in two-sided platforms: An analysis of bias. *Management Science* 68, 10 (2022), 7069–7089.

[13] Maxwell Kellogg, Magne Mogstad, Guillaume A Pouliot, and Alexander Torgovitsky. 2021. Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American statistical association* 116, 536 (2021), 1804–1816.

[14] Noémi Kreif, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. 2016. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics* 25, 12 (2016), 1514–1528.

[15] Kathleen T Li. 2024. Frontiers: A simple forward difference-in-differences method. *Marketing Science* 43, 2 (2024), 267–279.

[16] Michael W Robbins, Jessica Saunders, and Beau Kilmer. 2017. A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *J. Amer. Statist. Assoc.* 112, 517 (2017), 109–126.

[17] Jonathan Roth and Pedro HC Sant'Anna. 2023. Efficient estimation for staggered rollout designs. *Journal of Political Economy Microeconomics* 1, 4 (2023), 669–709.

[18] Liyang Sun, Eli Ben-Michael, and Avi Feller. 2025. Using multiple outcomes to improve the synthetic control method. *Review of Economics and Statistics* (2025), 1–29.

[19] Ruoxuan Xiong, Alex Chin, and Sean J Taylor. 2024. Data-driven switchback experiments: Theoretical tradeoffs and empirical Bayes designs. *arXiv preprint arXiv:2406.06768* (2024).