# Designing Experiments with Synthetic Controls

Nick Doudchenko David Gilinson Sean Taylor Nils Wernerfelt*

**Abstract**

Synthetic controls have become a powerful and standard component of the applied researcher's toolkit. Research to date often takes the treated unit as fixed and conducts post-hoc analyses of different interventions. A common problem that has become increasingly relevant in applied work, however, is given a set of possible test units, how can a researcher select the best one(s) to experiment on? This paper develops an approach for answering this question with synthetic controls, leveraging simulated interventions and permutation tests across candidate test units. We also discuss frequent implementation issues that may arise in practice and how they can be addressed. Finally, using historical data from Facebook, we demonstrate the design and analysis of a country-level experiment and show the substantial gains from utilizing this approach. Our methodology is implemented in the open source R package `countrytestr`.

**Keywords:** Power calculation, permutation tests, observational studies

*Nick Doudchenko: AI Resident, Google, 111 8th Ave, New York, NY 10011 (e-mail: nikolayd@google.com); David Gilinson: Data Scientist, Facebook, 1 Facebook Way, Menlo Park, CA 94025 (e-mail: gilinson@fb.com); Sean Taylor: Research Scientist Manager, Lyft, 185 Berry St #5000, San Francisco, CA 94107 (e-mail: sjt@lyft.com); Nils Wernerfelt: Research Scientist, Economics, Calibra (Facebook), 1 Facebook Way, Menlo Park, CA 94025 (e-mail: nilsw@fb.com). The authors would like to especially thank Guido Imbens for valuable discussions and feedback. We are also grateful to seminar participants at several tech companies and the Conference on Synthetic Controls and Related Methods at MIT. All remaining errors are our own.

# 1 Introduction

Since their inception, synthetic controls have become a standard component of the applied econometrician's toolkit. Given that many policy interventions of interest occur in a single location and at an aggregate level – circumstances well suited to synthetic controls – the methodology has been used to study to a wide range of applications. Analyses include the canonical examples of evaluating the impact of terrorism in the Basque region and California's 1989 cigarette tax to more modern examples of gun control, natural disasters, and the minimum wage [1, 2, 3, 4, 5]. Importantly, these numerous studies usually take the treatment unit as fixed and perform an ex post analysis. An increasingly common problem faced in applied work, however, is given a set of possible test regions, how can a researcher select which one(s) to experiment on?

One common reason researchers experiment at an aggregate level is simply exogenous constraints. For example, to evaluate the impact of a television ad campaign on purchasing behavior, a researcher may want to run an individual level experiment; since television ads in the US can only be purchased at the DMA level however, the question may then become which DMAs to launch in in order to best assess the ad's impact? As another example, when a company is testing a new product, supply chain constraints may mean that the product can only be able to be introduced at the region level. A researcher seeking to evaluate what effect this new product is having on existing sales may then have to select which regions to activate the relevant supply chains.

Another common reason for experimenting at an aggregate level is concerns over interference. Indeed, even when treatment at more granular levels is possible, researchers may prefer an aggregate intervention in such cases. For example, suppose a ride-sharing company were interested in understanding the effect of higher wages on driver behavior. One approach would be an A/B test that randomly boosted some drivers' wages; however, this would not capture the equilibrium effects of every driver within a market receiving higher wages and the resulting downstream effects on total ridership, competition, etc. For this reason, it may make more sense to experiment at the market level.

From a practical standpoint, exogenous constraints and concerns over interference drive a large number of experiments to happen at an aggregate level. For example, nearly every major company launches region-level tests today; the companies represented by the authors have launched hundreds alone. From an analysis perspective, if the number of markets is large and the costs per intervention are low, randomization at the region level may be useful in these settings. Very frequently, however, such large scale interventions are quite costly, meaning that only a few markets may be able to get the treatment and the researcher has to select which ones. In such settings, synthetic controls have become a powerful and attractive tool.

Consider now the problem from the vantage point of the experimenter: if she has to select which unit(s) get treated, how should she choose? Intuitively, if the aim of the experiment is to assess the intervention for a potential global launch, the selected units should be representative of the broader population. In

addition, the selected units should be such that if the intervention does affect the outcome of interest at the aggregate level, the experiment would maximize the probability of detecting it; similarly, if the intervention has no effect, the experiment should minimize the probability of detecting one. We note in classic synthetic control applications, researchers have not had to worry about selecting the treatment units and, by virtue of conducting ex post analyses, concerns about external validity, power, and bias have often been relegated to hopeful assumptions. If one can select the units though, these concepts become very important.

Our paper develops a method for designing experiments with synthetic controls to address these concerns. Our main innovation is in estimating the power for each candidate region given a metric and post-launch horizon of interest. Specifically, we leverage simulated interventions of different magnitudes across historical data within each potential test unit to estimate a power curve for each one. This represents a data-driven way to compare individual units according to their historical performance in detecting interventions of different sizes. In the case of multiple test units, since the covariance across units becomes important, our problem becomes one of optimal subset selection. Here we define an objective function and leverage a simulated annealing algorithm over the space of country subsets to further optimize our unit selection.

We then illustrate the methodology in practice, leveraging data from a historical country test from Facebook. Specifically, we walk through the unit selection problem, the post-launch analysis, and the robustness checks. These steps are done using the R package `countrytestr`, which is written with a specific implementation of our approach and is being released in conjunction with this paper. To the best of our knowledge, it is the only package that can support the full life cycle of a synthetic control experiment.

Finally, we provide evidence that the gains from leveraging our approach are substantial. To do this, we considered the simple case of selecting a single test country where we wanted to detect a multiplicative treatment effect that is homogenous across countries. Given a standard Facebook metric and common experimental horizon (14 days), selecting a country based on the estimated MSE from our approach versus one at random reduced the size of the minimum detectable effect by 41% and the bias by 75%. In addition, close to 20% of the countries had false positive rates in excess of 5% (with some as large as 50%), which we interpreted as a cautionary note against analyses where the experimenter has to take the treated unit as a given.

Taken together, we view our approach as a very practical method for doing observational studies in synthetic control settings. In the spirit of [6], we view the design of such observational studies as an important and understudied area in between pure experimental design and observational study analysis.[1] We hope that our framework and accompanying R package will improve the use and efficacy of synthetic controls even further.

The rest of this paper is organized as follows. Section 2 provides more detail on the synthetic control

[1] See [7] for another example of a recent paper in this tradition.

methodology; Section 3 discusses our approach for unit selection; Section 4 walks through an example leveraging our R package and data from the design and analysis of a country-level experiment at Facebook; Section 5 analyzes how much is gained by leveraging our approach; and finally Section 6 concludes.

## 2 Synthetic Controls

The synthetic control methodology, introduced in [1], creates a counterfactual prediction for trends in a treated unit based on a combination of pre-treatment data from both the control and the treatment units, and post-treatment data from the control units. Detailed descriptions of the synthetic control methodology can be found in [2] and [8]; we refer the interested reader to those papers and rather here present a broad synopsis.[2]

Following the notation in [9], suppose we observe data for $J+1$ units in periods $t = 1, 2, ..., T$. Abstracting away from the need to select which unit to test in for now, suppose unit 1 is treated during time periods $T_0 + 1, ..., T$. The remaining $J$ units are all untreated and constitute potential controls. Let $Y_{it}^N$ be the outcome for unit $i$ at time $t$ in the absence of the treatment and let $Y_{it}^I$ denote the outcome that would be observed if that unit were exposed to the intervention in those periods.[3] The aim of the synthetic control methodology is to estimate for $t > T_0$ the effect of the intervention on the treated unit:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N, \tag{1}$$

where the last equality follows from the fact that $Y_{1t}^I = Y_{1t}$, i.e., we observe only the treated state of the world for unit $i$. Let $\mathbf{W} = (w_2, ..., w_{J+1})'$ denote a set of weights for the metrics in the possible control units. Our estimate of the treatment effect is given by:

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}. \tag{2}$$

From this core structure, the literature has spread out in several directions. One large area of ongoing research is different methodologies for defining the weights, $w_j$. For instance, while early work imposed that the weights sum to one and be non-negative, more recent work has explored not only weakening those assumptions but also estimating the weights in the presence of complexities such as missing data, imperfect pre-treatment fit, or fixed effects (e.g., [8, 10, 11, 12]). Apart from methods for deriving the weights, researchers have also been interested in inference in synthetic control settings. On this topic, from the original synthetic control paper that had no confidence intervals, the literature has progressed to a much

[2] From an implementation perspective, see also [9] for details on the `Synth` R package.

[3] Note that our approach will ultimately allow us to analyze the average per period effect between launch and some time $t$ as well as the specific effect at time $t$. For now, assume this is referring to the latter.

more formal understanding of the assumptions and methods needed for proper inference (e.g., [13, 14]).

We build on top of these two literatures, though our paper has a different focus. Specifically, instead of optimizing features of the synthetic control itself, we aim to optimize the overall observational study conditional on a preferred synthetic control methodology. We now discuss our methodology in more detail, highlighting both the connection to past synthetic control work as well as to experimental design.

# 3 Unit Selection Methodology

There are several steps a researcher must consider in selecting test units. In this section, we cover them in chronological order for a researcher, starting from a method for calculating the synthetic control weights and confidence intervals, to calculating power and Type 1 error rates (for a one-region or multi-region test), weighing external validity, and finally considering other, pragmatic concerns that may arise.

## 3.1 Generation of Point Estimates and Confidence Intervals

In selecting possible units to experiment on, the first item the researcher needs is a favored methodology for estimating the synthetic control weights $w_j$. For expositional purposes, we present the specific implementation that we use in our R package, and we note that [14] provide conditions under which our inference procedures are valid.

Following other papers (e.g., [15, 16, 17]), we generate the weights via Lasso [18]. Specifically, if $y$ are the values of the metric of interest in the potential test unit across $t = 1, 2, ..., T_0$ periods in the training period with $p$ potential control units with corresponding metrics $x$, we solve

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{t=1}^{T_0} \left( y_t - \beta_0 - \sum_{j=1}^{p} x_{tj}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

In our R package, we somewhat arbitrarily select the model from Lasso where we restrict the maximum number of coefficients to be 12. This number was chosen as empirically it performed well with internal Facebook data, but more importantly our procedure runs a number of computationally expensive calculations to both conduct power calculations and analyze tests, so being able to solve quickly for the weights is essential.

Given this procedure to create point estimates, we generate the confidence intervals via permutation over past dates. Specifically, for a researcher-specified training window, we retrain Lasso (and, as a result, the controls and weights for a given test unit will vary across each iteration) starting at all possible historical dates, and then compare the predicted value against the observed data at each horizon. This generates an empirical distribution of residuals at each horizon, which is then used to produce confidence intervals. Such permutation tests are standard ways to generate confidence sets for synthetic controls, and [14] provide formal conditions under which confidence intervals generated in this manner will be exact.

## 3.2 Power Curve Calculation Details

The earlier section describes how we generate the point estimates and confidence intervals for a given test unit. In this section we describe how to take those point estimates and confidence intervals and map them into power curves for each candidate unit. There are two methods we developed, which differ if a researcher is interested in experimenting in a single region or multiple regions; we talk through each below.

### 3.2.1 One Region Case

Suppose a researcher wants to find the one best region to experiment in from a power perspective. The intuition for our approach can be seen in Figure 1. Namely, suppose we have a candidate region and we are interested in evaluating what power we will have for a given post-launch horizon and metric of interest.[4] We first take the historical time series data of that metric across time in the candidate region. We then select a random date and take all data after that day and shift it up by a specified effect size (Figure 1b). For example, in this figure, all data after $t = 550$ is shifted up by 20%. We then run our synthetic control procedure on the data, generating a forecasted counterfactual and confidence interval (Figure 1c). If the shifted data falls outside the confidence interval, then we flag that this effect size was detected in this potential test unit. We repeat this procedure many times – over different dates and effect sizes – and use the average empirical rejection rates of the null hypothesis to generate both a power curve and a reported Type 1 Error Rate.[5]

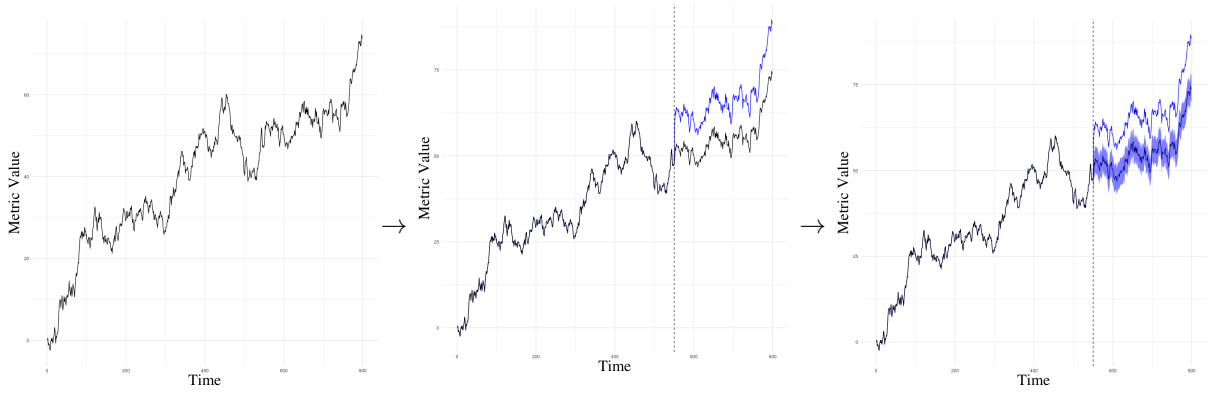
Two other notes on the procedure. First, if the researcher believes the true effect is not modeled as a level shift but instead some other functional form (e.g. a trend shift), the procedure can be reproduced assuming such an intervention.[6] Second, it is worth noting that as part of this procedure, the algorithm outlined in the previous section is repeated many times for each potential test unit and it is essential that that procedure is optimized to perform those calculations and permutations rapidly.

In conclusion, for a metric and horizon of interest, this approach can be used to empirically rank regions by the power and Type 1 error rates they will have. This is useful if we want to launch in a single region and believe the true effect is homogenous. Often, however, we may have the bandwidth to test in multiple regions or think effects may be heterogenous. We explore such issues next.

[4]The horizon may be the entire post-launch window, or subperiods. Empirically, we have found many of the country tests at Facebook have required a burn in period for novelty effects to wear off. In such cases, it may be best to set the horizon of interest to, for example, the window between 21 and 51 days post-launch. To the best of our knowledge, no other R package implements functionality for analysis of such subperiods.

[5]At least at tech companies, if an intervention is being considering at a given geographic level, empirically it is quite common to have lots of historical data stored that is aggregated to the proper level. This is because many interventions occur at aggregation levels that are commonly of interest for all sorts of other purposes (e.g. advertising, metric monitoring, etc.).

[6]Currently, our R package only supports multiplicative interventions – e.g., a 5% increase in a metric, as that is empirically what there has been the most demand for.




(a) Raw (simulated) data.

(b) Artificially shifted data after fake intervention

(c) Analysis with synthetic control

Figure 1: The power calculation process starts with the raw historical data (simulated AR(1) in this case) as in the first subfigure. As described in the text, we then launch many simulated fake interventions of different magnitudes and across different dates, and leverage the empirical null hypothesis rejection rate to generate a power curve and Type 1 Error Rate.

### 3.2.2 Multi-Region Case

Often, one may wish to experiment in multiple regions. For example, to increase power or study potential heterogenous effects. When combining treatment effect estimates across regions, the combined variance must take into account the covariance across regions. From an ex ante perspective, this makes selection of test regions an optimal *subset* selection problem, which is computationally very involved.

Note that in practice, analyzing each country individually and then estimating the combined power still yields valid results; the point here is that there might exist another subset with better experimental properties. To try to estimate that subset, we devise a procedure based on simulated annealing, and implemented in our package as `experimental.unit.set.search`. This procedure works as follows.

First, choose a 'temperature' $T_0 > 0$, a 'cooling factor' $\kappa < 1$, and an objective function $f$ (e.g. the minimal detectable effect obtained in a power analysis for a set of countries). Start with an initial guess of the set of test units, $\mathcal{S}_0^{test}$ and set the initial value of the objective function at $f_{-1} = \infty$. At iteration number $j = 0, 1, 2, ....$ evaluate the objective function $f_j$ and compute the improvement in the value of the objective, $\delta_j \equiv f_j - f_{j-1}$. Then, compute the probability of switching to a candidate test set

$$p(\delta_j, T_j) = \begin{cases} 1, & \delta_j \leq 0 \\ \exp\left(-\delta_j / T_j\right), & \delta_j > 0 \end{cases}$$

After this, set the new temperature, $T_{j+1} = \kappa T_j$ and propose a new candidate test set, $\mathcal{S}_{j+1}^{test}$, randomly replacing one unit. Replace this procedure until the maximum number of iterations is reached.

In our implementation of this package, we throw out countries that have high Type 1 Error Rates and set the objective function to minimize as the width of the power curve at 80%. This is due to the fact that empirically the kinds of interventions we have studied are better to estimate imprecisely than with bias. However, any objective function could be plugged into this simulated annealing process to be maximized (e.g. MSE, minimum detectable effect, etc.).

### 3.3 External Validity

The earlier Subsection 3.2.1 was focused on optimizing for power. If we assume a homogenous treatment effect across countries, that approach will help us find an experimental set optimized along that dimension. However, experimenters often have reason to believe effects are heterogenous across regions, and they want to get an estimate of the global average treatment effect that would occur if the product launched across their entire domain.

Leveraging standard experimental design results, there are three potential answers here we have seen discussed in practice. First, the time of launch and country itself can be selected randomly. Often exogenous constraints make this difficult, but in theory that would provide valid inference for the question of a global launch. Two comments on this approach though. First, empirically we have found massive heterogeneity in confidence interval size across candidate regions. Hence, choosing a region purely at random, while reducing the bias, likely has a massive variance effect, making it suboptimal from an MSE perspective. It is up to the experimenter to make this tradeoff, and the information provided by this analysis can only lead to a more informed decision. Second, empirically we are clearly limited in terms of choosing the timing at random, and it may well be worth it to run the same experiment at different time periods to understand if there is any heterogeneity across time.

Second, if the researcher believes she knows observables along which effects are heterogenous, she can optimize within each strata and then reweigh accordingly for the combined average treatment effects. For example, an experimenter at a ride-sharing company may believe that the effect of an intervention may vary by if a market is in a rainy location or sunny location. She can then launch experiments in each kind of market and weigh according to the share of markets that are rainy versus sunny to estimate the combined treatment effect.

Third, the experimenter could do partial randomization, whereby the experimenter first limits the set of possible countries down, and then randomizes within them. Pragmatically, we have found there are almost always exogenous constraints (discussed later) that ex ante put a substantial fraction of the world off limits for the experimenter, thereby already being a threat to estimating a global average treatment effect. If the experimenter feels the treatment effect does not correlate with the power too substantially, then doing a first pass screen based on power and a randomization within that set may be optimal from an MSE perspective.

### 3.4 Pragmatic Constraints

Finally, we wanted to comment on some pragmatic issues that we have seen arise in designing such experiments in practice. In particular, there are three kinds of such constraints we have seen come up in empirically designing these tests: (i) some regions may be off limits to experiment in, (ii) SUTVA violations caused by underlying market structure, and (iii) time shocks – either historical or in the future – that are outside of the experimenter's control. We will now briefly discuss each of these.

In practice, a common issue that may come up is that certain regions may not be open to the experimenter. For example, legal, political, or institutional reasons may preclude a country from being a viable candidate. This may result in threats to the external validity of the experiment that the researcher will not be able to mitigate. In practice, we have found this often to be a cost-benefit analysis at the firm level. Specifically, the firm will has to weigh the pros/cons of the increased experimental benefit versus the potential harm to the other firm-related outcome variables.

Another source of exogenous constraints we have seen are SUTVA violations, whereby, for example, activities within two markets may affect each other and testing in one may affect the other which may be a potential control. This concern is very market and metric-specific. Videos produced in one country may be heavily watched by users in another, for example. Empirically, we have mitigated this concern by looking at our metrics of interest, and excluding as potential controls markets that we believe are likely contaminated. Since the synthetic control takes a weighted combination of so many countries, in country tests at Facebook at least we have found excluding such markets to have a minimal effect on power.

Finally, we have found idiosyncratic, region specific shocks to be a threat to experimental validity. For example, launching experiments during Ramadan, holidays, elections, or unexpected news cycles all present possible threats. In practice, we have found that the best an experimenter can do is research possible future events in different candidate regions, check historical data for such anomalies, and then drop outliers, winsorize, use moving averages, or remove candidate test or control units as appropriate.

## 4 Example Application with Facebook Data

Given this background, we now demonstrate the implementation of this methodology leveraging our R package `countrytestr` with Facebook data.[7] Specifically, we walk through the life cycle of a region-level test from the perspective of a data scientist who wants to select which countries to test in, analyze the experiment post-launch, and then perform robustness checks.

[7]Note, to preserve privacy, the names of the countries and dates have been shuffled, as has the launch date of this test. The metric has also been rescaled from its original value.

## 4.1 Set up

For simplicity, we will walk through an example that focuses on one metric at one horizon of interest.[8]

First, we load the package and display our sample data set, the data table `wide.dt`. This data table contains panel data at the day level for many countries in the world for a certain Facebook metric. The data is sorted by day ('ds') and goes back until June 2015; importantly, we have numeric values for every entry of our data - there are no missing, NA, or infinite values.[9]

```
> library("countrytestr")
> head(wide.dt)
  ds           HN       MN       TT       CU        TM       US
1: 2015-06-25 70306.04 98937.17 112888.0 86012.78 84985.77 92429.23
2: 2015-06-26 70268.40 99014.96 113040.0 85892.35 85263.25 93036.02
3: 2015-06-27 70835.41 99032.06 113018.4 85152.63 85061.10 93092.16
4: 2015-06-28 70693.07 99197.21 112910.1 84415.24 86235.18 92779.76
...
```

Any dataset that gets put into `countrytestr` should be of this format. There is also no need for standardization, as our package leverages `glmnet` to run Lasso, which standardizes the variables for us.

## 4.2 Power Calculation

Given the above data, the data scientist can now compare different countries to see which have the smallest bias and confidence interval size at the horizon of interest. To run such an analysis for a specific country, the researcher leverages the `power.analysis` function and simply enters:

```
pa.results <- power.analysis(
  data = wide.dt[ds < launch.date],
  n.simulated.launch.dates = 30,
  max.horizon = 21,
  test.units = c('CM'),
  effect.sizes = -10:10 / 100
)
```

[8] If an experimenter cares about multiple metrics over multiple time horizons, it is straightforward to generalize this example to those settings by repeating this procedure across metrics and horizons and then making decisions based on the collective results.

[9] If such issues arise, we recommend using interpolations or only starting the historical data after problematic stretches.

Let us break down what each of these arguments does:

- `data`: this is a data table of the form shown above. Here we are using data that include both pre- and post- launch, we restrict to data prior to the launch date to mirror the information set we would have in designing the experiment.

- `max.horizon`: the horizon we are interested in observing an effect, in days – in other words, this says we want to know three weeks after launch what our power will be for the average effect from launch to day 21. To run the power calculation for a different window (e.g. a specific day or week post-launch), the user can input a `min.horizon` as well.

- `n.simulated.launch.dates`: The number of dates to use in the power simulation. Specifically, the `power.analysis` function will take `n.simulated.launch.dates` different possible start dates, run fake interventions of different sizes at each of those dates, and detect what power the synthetic control would have at the horizon of interest. It leverages the most recent possible data for this (i.e., starting at `max(wide.dt[ds < launch.date]$ds)- max.horizon` and including the previous `n.simulated.launch.dates` days).[10]

- `test.units`: the specific country we are interested in evaluating. In this case, Cameroon ('CM'), though combinations of countries are also supported – e.g. `test.units = c('CM','US')` would give the power from a test launched in Cameroon and the US. More details later on how the package aggregates effects across multiple test units.

- `effect.sizes`: the sizes of interventions we want to test. This simulates increases of (in the above example) -10%, -9%, ...+10% and then sees if we have the power to detect each effect size. The output is a power curve (shown below) which thus tells us which size intervention we'll be powered to detect. More possible effect sizes take longer but also give more precise output since we interpolate between entered effects to estimate the exact power.[11]

This will run for a few minutes - the runtime is largely determined by the number of effect sizes and simulated launch dates. Once it is done, we can see the output as follows:

```
> pa.results
```

[10] We commonly use between 30 and 90 days; if the answers are very different between these, that suggests there may be a lack of stationarity in the data, which is concerning.

[11] As mentioned earlier, the package currently only supports multiplicative interventions. If one were interested in additive or other functional forms, though, one could manually recode the `power.analysis` function to support those specifications.

```
Power analysis run for 30 days between 2017-10-26 and 2017-11-24
Test of cumulative ATE from horizon 0 to 31 with 95% confidence has 80% power for effect sizes
    (interpolated):
Lower: -5% (-4.545%)
Upper: 5% (4.429%)
Type I Error Rate: 0%. This is OK compared to the acceptable rate of 5%
```

The interpretation is relatively straightforward. Namely, over the first 21 days if average daily effect size is greater than or equal to 5%, the 95% confidence interval derived from the `countrytestr` package will be able to detect it 80% of the time.[12] Note that due to the permutation test, there is no guarantee the upper and lower envelopes of the power curve are symmetric, and so the number is slightly different for a negative effect. If instead of finding the effect averaged over the entire time horizon, the experimenter would prefer the effect over a specified window (e.g. one day or a specific week post-launch, for example, to avoid novelty effects), she can specify a `min.horizon` in the power calculation to obtain results for the appropriate window.

As discussed earlier, the Type I error rate comes from launching fake interventions of different sizes on historical data and reporting the empirical rejection rates for the 95% confidence interval at the horizon of interest. If this number is above 5%, the experimenter should be worried that the proposed country-metric pair is biased at that horizon and consider alternatives.

For more detailed results, we can plot the overall power curve to tell us at all `effect.sizes` the estimated power `countrytestr` will have. Note that bias in the historical predictions will manifest in an off-center curve; in the example here, there appears to be no evidence of bias.

```
> plot(pa.analysis)
```

[12] The 'cumulative' comes from averaging the effect over the whole horizon.

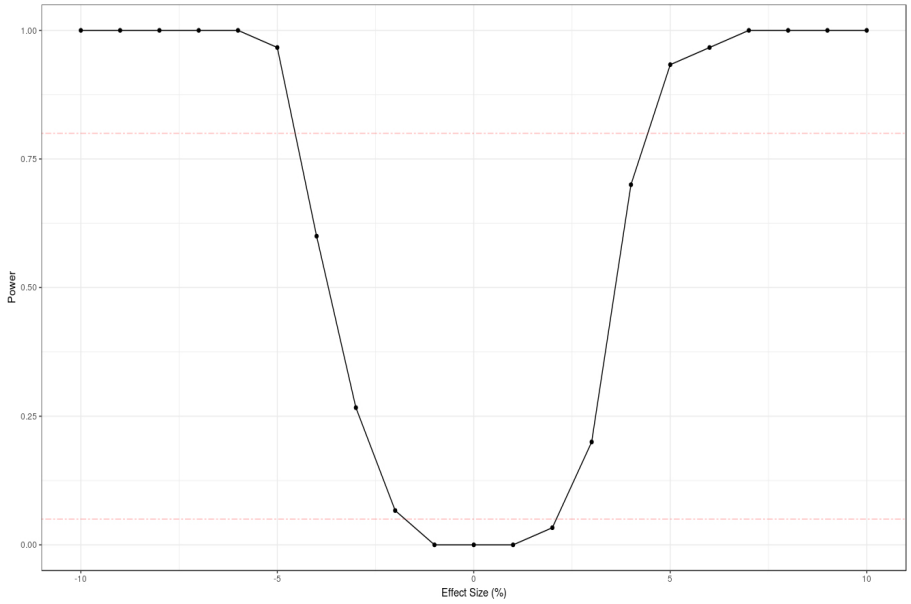



Figure 2: Power curve for the metric of interest at the horizon (21 days) and potential test country (CM).

It is worth caveating these numbers by saying these are based on historical data and permutation tests in such settings are only valid under conditions laid out in [14]. Pragmatically, if the experimenter reruns `power.analysis` with different `n.simulated.launch.dates` and the power curves are relatively stable, that helps reassure the experimenter that at least historically, a power curve derived on one day is likely similar to what she can expect from a launch soon thereafter.

As with any analysis involving synthetic controls, it is important to also graph the trends in test and control units to see if anything suspicious is happening that might violate the parallel trends assumption. We note that due to different time varying trends in countries, the control units (and weights) for a given country on one day will not necessarily equal those on another day. Empirically, we recommend taking the most recent data, and plotting for each potential test country the trends in it and the control units. Problematic regions and/or time periods can be omitted or smoothed over as the experimenter decides.[13] We discuss in the post-launch analysis section how to create such plots using the `control.plot` function.

To summarize this section, given an input matrix of daily data for many regions, the `power.analysis` function can provide estimates of the confidence interval size and bias an experimenter can expect for a given potential test region, metric, and horizon. The function is straightforward to adjust to explore many potential variants of these inputs as well. Finally, we note that our example uses day level data, but the same analysis could be done using data aggregated over any time window, provided enough data from either the past or other regions is available.

[13] In practice, at least for tests at the country level, given the number of possible control units, we have found that dropping some from the data affects power minimally.

### 4.3 Analyzing the experiment

After selecting the test unit and launching the test, the question at hand then becomes how to analyze the experiment. After launch, the analysis is straightforward to run with `countrytestr`:

```
results <- countrytestr(
  wide.dt,
  launch.ds=as.Date(launch.date),
  test.units=c('CM'),
  max.horizon=31
)
> results
Cumulative Average Estimated Treatment Effects for 2017-12-26 thru 2018-01-26:
Average Estimated Treatment Effect: -13.7% [-17.3%, -10.2%] with 95% confidence.
There is a 0% chance of observing an effect this large or larger assuming treatment effect is zero.
```

The terms here should be straightforward based on the past discussion, but we take in all available data (pre- and post- launch), and given a launch date, return the output. We note as well that to compute the effect over a different time window, the experimenter simply has to add a `min.horizon` above. As predicted, the post-launch confidence interval size is consistent with that predicted by the power calculation, and the intervention was sufficiently powerful so as to yield a significantly negative effect.

We can further visualize the results graphically across time using `> plot(results)`, which returns:

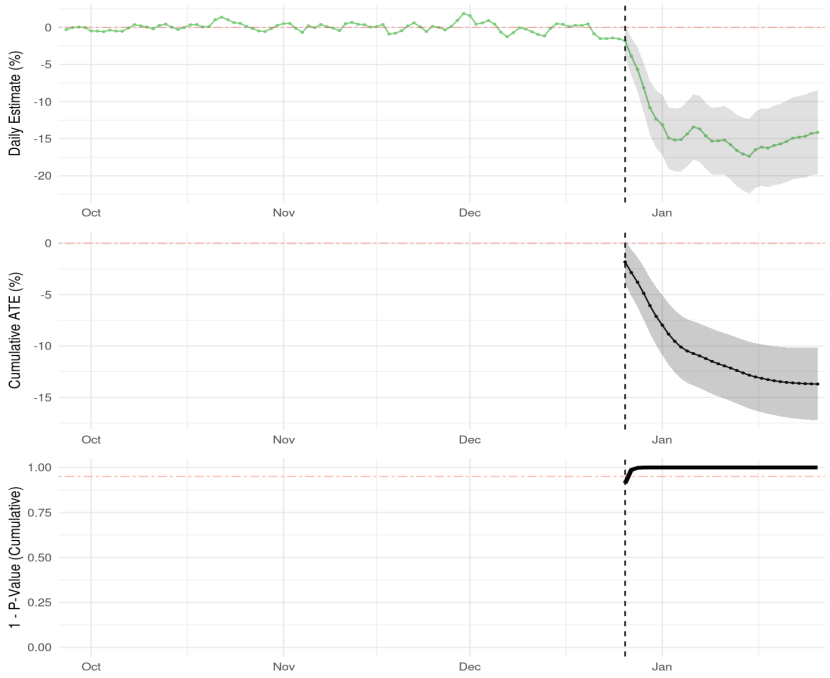



Figure 3: Plot of treatment effect across time; 95% confidence interval is in green, with the launch date denoted by the vertical dashed line. The top graph shows the (recentered) difference between the synthetic control and the trend in the test country pre-launch, followed by the estimate effect size and confidence interval. The middle graph shows the cumulative average effect across all days plotted, and the last is the $p$-value for the cumulative effect.

To see only the plot for the raw data in the test country, we can leverage `> unit.plot(results)`. In the event that there any multiple test countries, this command will show them in parallel:

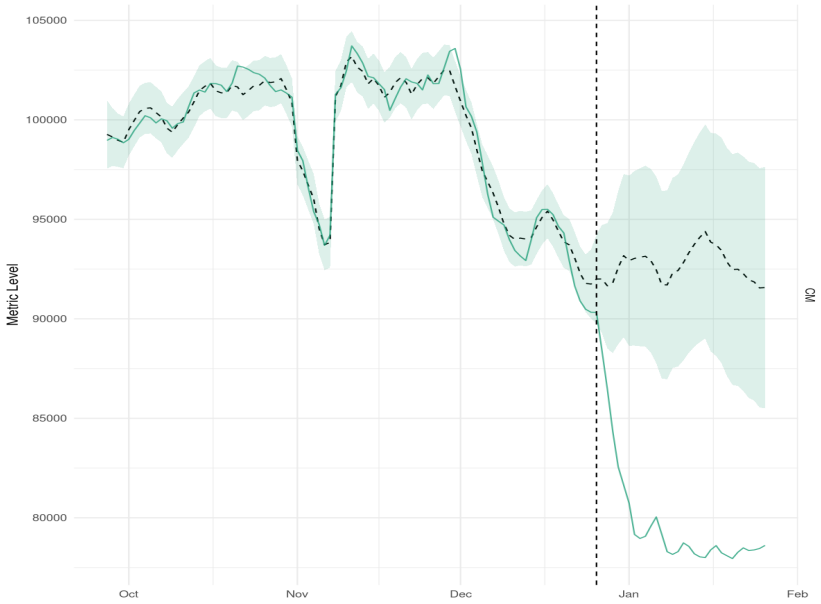



Figure 4: Raw data in test country, with synthetic control pre- and post- launch as well as empirical trend.

To see what countries have been selected as controls (and what their weights are) for each test country, we can type the below, which will also spit out the weights generated by the Lasso run at launch day.

```
> summary(results)
Control Unit Weights:
CM: RO [30.3%], SO [13.85%], KE [8.94%], MN [8.82%], MS [8.2%], ZM [7.18%], AO [7.14%], GR
    [6.75%], NL [4.32%], FR [3.15%], ML [1.35%]
```

Finally, for a given test unit, to see the trends in these control countries, `> control.plot(results, test.unit = 'CM')` returns a graph of all those trends. This helps provide assurance on the parallel trends assumption. As aforementioned, this command can also be used during the power analysis to see for a potential test country, what the current estimation of the control countries and weights is.

## 4.4 Multiple Countries

The above example considers the problem of selecting only one country. `power.analysis` can easily estimate the power for a combination of countries, by changing the `test.units` to list multiple countries (e.g.

`test.units=c('CM','US')`). We will now talk about how `power.analysis` combines the multiple test units and how that can increase power.

To combine the two estimates, `power.analysis` treats each country as an equally weighted observation. The combined average treatment effect is the average of that from the treated units, and the variance is given by their sums, including the covariance terms. This means that experimenting in additional countries can help increase the precision of the combined average treatment effect. How many countries should one experiment in then, and how should they be selected?

In terms of number of countries, we believe synthetic control experiments are most useful when we have a small number of test units. For example, if the option exists to experiment on half of all countries in the world, then a matching analysis or pure randomization may both provide better power and more reliable estimates. A practical reason for this is that for a fixed number of countries, as the number of test units gets large, there are both fewer possible control countries and if some countries are serving as controls for multiple test units, that increases the risk of a correlated shock. Whether matching is preferred or not depends upon the assumptions the experimenter thinks are likely to hold; we note though, that this can also be approached empirically, with evaluating matching techniques akin to what we are doing in `power.analysis` – apply to several different historical dates and see what the empirical power has been.

Working with a small number of test units makes the experimenter's job easier, but given the combined estimate takes into account the covariance across units, the problem now becomes one of choosing an optimal *subset* from all possible candidates. This is a much harder question and why we built `experimental.unit.set.search` (see earlier section on this topic). Using that tool, as well as prior experimenter knowledge, one can select optimal subsets to launch in from a power perspective.

# 5 How much does this approach buy us?

To evaluate how much experimenters can gain from our procedure, we consider the simple case of designing a single country experiment with a standard Facebook metric at a common experimental horizon of interest (14 days). As implemented in our R package, we analyzed a multiplicative treatment effect that is homogenous across countries, and then ran `countrytestr` over each of 101 possible countries. To be clear, we ran the `power.analysis` function for each country and pulled out the false positive rate (FPR), the interpolated upper and lower bounds of the power curve, and the average bias from that command. We can then compare across all countries based on these metrics and select according to whatever criteria we would like.

The results of this exercise are in Table 1. Note that the entries to the left of the double vertical lines represent the columns statistics across all countries (e.g., the average FPR across all countries, the min FPR across all countries, the average value of the absolute value of the bias across all countries, etc.); the entries to the right of the double vertical line refer to specific countries (i.e., the FPR, Abs(Bias), and MDE for the

country with the minimum MSE value, etc).

Several observations from this table. First, note that the expected MDE and bias from selecting a random country are 6.13% and 1.35%, respectively. In contrast, selecting the country with the lowest MSE has MDE and bias of 3.63% and 0.34%. (The FPR also decreases from 4.22% to 0%.) Hence, leveraging this approach in this example, we can reduce the size of the MDE by 41% and the bias by 75%.

Second, from this we can see that some countries have very poor properties, even for the standard, commonly used metric. For example, a false positive rate of 50% or a bias of 10%. The MDE of 10% is due to the fact that this analysis was run with effect sizes ranging from -10% to +10%, meaning that MDE's greater than 10% in absolute value were entered as NA values by `power.analysis`. To compute the numbers for this table, the NA's were replaced with ±10%, but note that this is means that the gains from optimizing based on MSE versus selecting a random country as reported here are actually *underestimates* of the true effect, as we do not observe all the MDE's that were larger in absolute value than 10%.

Third, the last two columns report other countries, chosen based on minimizing the MDE and bias for all countries with a reasonable FPR. We note that these also outperform selecting a country at random.

Separately, though it is not in the table, 19/101 countries had FPR's greater than 5%. In other words, selecting a country at random to experiment in will yield a country with an overly high FPR about 20% of the time. This is of course conditional on our method of analysis, but we think still is informative about risks from synthetic control analyses where the unit is exogenously set.

| | Average | Min | Max | Min(MSE) | Min(MDE \| FPR $\leq$ 5%) | Min(Abs(Bias) \| FPR $\leq$ 5%) |
|---|---|---|---|---|---|---|
| FPR | 0.0422 | 0 | 0.5 | 0 | 0 | 0 |
| Abs(Bias) | 0.0135 | 0.0001 | 0.1001 | 0.0034 | 0.0034 | 0.0001 |
| MDE | 0.0613 | .031 | 0.1 | 0.0363 | 0.0325 | 0.0567 |

Table 1: Analysis of False Positive Rate (FPR), bias, and Minimum Detectable Effects (MDE) at 80% power across 101 countries with Facebook data. To the left of the double vertical line the row entries refer to their column values across all countries (e.g. across all countries, the max FPR is 0.5; across all countries the max bias is 0.1001 - or 10%, etc.); the columns on the right refer to specific values based on countries that, for example, have the lowest MSE. The MDE was calculated by taking the average of the absolute value of the left and right interpolated effects from `power.analysis`.

# 6 Conclusion

Though much early work in synthetic controls focused on ex post analyses, the question how to select the units to be treated for a synthetic control analysis has become increasingly popular. We provide a

methodology that can be used to guide that decision and demonstrate that it can engender substantial gains on the experimenter's side. Further, we provide an R package implementation that can support the full life cycle of a synthetic control experiment, as well as an example of it using historical Facebook data.

# References


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

[2] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 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, 2010.

[3] John J Donohue, Abhay Aneja, and Kyle D Weber. Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. Technical report, National Bureau of Economic Research, 2017.

[4] Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano. Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, 95(5):1549–1561, 2013.

[5] Ekaterina Jardim, Mark C Long, Robert Plotnick, Emma Van Inwegen, Jacob Vigdor, and Hilary Wething. Minimum wage increases, wages, and low-wage employment: Evidence from seattle. Technical report, National Bureau of Economic Research, 2017.

[6] Donald B Rubin et al. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

[7] Ruoxuan Xiong, Susan Athey, Mohsen Bayati, and Guido W Imbens. Optimal experimental design for staggered rollouts. *Available at SSRN*, 2019.

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

[9] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13), 2011.

[10] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

[11] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*, 2018.

[12] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

[13] Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, pages 1–16, 2019.

[14] Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*, 2017.

[15] Kathleen T Li and David R Bell. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65–75, 2017.

[16] Rafael Valero. Synthetic control method versus standard statistic techniques a comparison for labor market reforms. 2015.

[17] Carlos Carvalho, Ricardo Masini, and Marcelo C Medeiros. Arco: an artificial counterfactual approach for high-dimensional panel time-series data. *Journal of econometrics*, 207(2):352–380, 2018.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.