Synthetic control

"I'm representin' for them gangstas all across the world"
– Dr Dre

> "The synthetic control approach developed by Abadie et al. [2010, 2015] and Abadie and Gardeazabal [2003] is arguably the most important innovation in the policy evaluation literature in the last 15 years." - Athey and Imbens [2017]

In qualitative case studies, such as de Toqueville's classic Democracy in America, the goal is to reason inductively about the causal effect of events or characteristics of a single unit on some outcome using logic and historical analysis. But it may not give a very satisfactory answer to these causal questions because oftentimes it lacks a counterfactual. As such, we are usually left with description and speculation about the causal pathways connecting various events to outcomes.

Quantitative comparative case studies are more explicitly causal designs. They usually are natural experiments and they usually are applied to only a single unit, such as a single school, firm, state or country. These kinds of quantitative comparative case studies compare the evolution of an aggregate outcome with either some single other outcome, or as is more oftentimes the case, a chosen set of similar units which serve as a control group.

As Athey and Imbens [2017] point out, one of the most important contributions to quantitative comparative case studies is the synthetic control model. The synthetic control model was developed in Abadie and Gardeazabal [2003] in a study of terrorism's effect on aggregate income which was then elaborated on in a more exhaustive treatment [Abadie et al., 2010]. Synthetic controls models optimally choose a set of weights which when applied to a group of corresponding units produce an optimally estimated counterfactual to the unit that received the treatment. This counterfactual, called the "synthetic unit", serves to outline what would have happened to the aggregate treated unit had the treatment never occurred. It is a powerful, yet surprisingly simple, generalization of the differences-in-differences strategy. We will discuss it now with a motivating example - the famous Mariel boatlift paper by Card [1990].

Cuba, Miami and the Mariel Boatlift

"Born in Miami, right on time
Scarface, El Mariel, Cuban crime"
– Pitbull

Labor economists have debated the effect of immigration on local labor market conditions for many years [Card and Peri, 2016]. Do inflows of immigrants depress wages and the employment of natives in local labor markets? For Card [1990], this was an empirical question, and he used a natural experiment to evaluate it.

In 1980, Fidel Castro announced that anyone wishing to leave Cuba could do so if they exited from Mariel by a certain date, called the Mariel Boatlift. The Mariel Boatlift was a mass exodus from Cuba's Mariel Harbor to the United States (primarily Miami Florida) between April and October 1980. Approximately 125,000 Cubans emigrated to Florida over this six month period of time. The emigration stopped only because Cuba and the US mutually agreed to end it. The event increased the Miami labor force by 7%, largely by depositing a record number of low skill workers into a relatively small area.

Card saw this as an ideal natural experiment. It was arguably an exogenous shift in the labor supply curve, which would allow him to determine if wages fell and employment increased, consistent with a simple competitive labor market model. He used individual-level data on unemployment from the CPS for Miami and chose four comparison cities (Atlanta, Los Angeles, Houston and Tampa-St. Petersburg). The choice of these four cities is delegated to a footnote in the paper wherein Card argues that they were similar based on demographics and economic conditions. Card estimated a simple DD model and found, surprisingly, no effect on wages or native unemployment. He argued that Miami's labor market was capable of absorbing the surge in labor supply because of similar surges two decades earlier.

The paper was very controversial, probably not so much because he attempted to answer empirically an important question in labor economics using a natural experiment, but rather because the result violated conventional wisdom. It would not be the last word on the subject, and I don't take a stand on this question; rather, I introduce it to highlight a few characteristics of the study.

It was a comparative case study which had certain strengths and weaknesses. The policy intervention occurred at an aggregate level, for which aggregate data was available. But the problems with the study were that the selection of the control group is ad hoc and ambiguous, and secondly, the standard errors reflect sampling variance as opposed to uncertainty about the ability of the control group to reproduce the counterfactual of interest.[138]

Abadie and Gardeazabal [2003] and Abadie et al. [2010] introduced the synthetic control estimator as a way of addressing both simultaneously. This method uses a weighted average of units in the donor pool to model the counterfactual. The method is based on the observation that, when the units of analysis are a few aggregate units, a combination of comparison units (the "synthetic control") often does a better job of reproducing characteristics of a treated unit than using a single comparison unit alone. The comparison unit, therefore, in this method is selected to be the weighted average of all comparison units that best resemble the characteristics of the treated unit(s) in the pre-treatment period.

Abadie et al. [2010] argue that this method has many distinct advantages over regression based methods. For one, the method precludes extrapolation. It uses instead interpolation, because the estimated causal effect is always based on a comparison between some outcome in a given year and a counterfactual in the same year. That is, its uses as its counterfactual a convex hull of control group units, and thus the counterfactual is based on where data actually is, as opposed to extrapolating beyond the support of the data which can occur in extreme situations with regression [King and Zeng, 2006].

A second advantage has to do with processing of the data. The construction of the counterfactual does not require access to the post-treatment outcomes during the design phase of the study, unlike regression. The advantage here is that it helps the researcher avoid "peaking" at the results while specifying the model. Care and honesty must still be used, as it's just as easy to also look at the outcomes during the design phase as it is to not, but the point is that it is hypothetically possible to focus just on design, and not estimation, with this method.

Another advantage, which is oftentimes a reason that people will object to the study ironically, is that the weights which are chosen make explicit what each unit is contributing the counterfactual. Now this is in many ways a strict advantage, except when it comes to defending those weights in a seminar. Because someone can see that Idaho is contributing 0.3 to your modeling of Florida, they are now able to argue that it's absurd to think Idaho is anything like Florida. But contrast this with regression, which also weights the data, but

[138] Interestingly, a recent study replicated Card's paper using synthetic control and found similar results. [Peri and Yasenov, 2018].

does so blindly. The only reason no one objects to what regression produces as a weight is that they *cannot see the weights*. They are implicit, rather than explicit. So I see this explicit production of weights as a distinct advantage because it makes synthetic control more transparent than regression based designs.

A fourth advantage, which I think is often unappreciated, is that it bridges a gap between qualitative and quantitative types. Qualitative researchers are often the very ones focused on describing a single unit, such as a country or a prison [Perkinson, 2010], in great detail. They are usually the experts on the histories surrounding those institutions. They are usually the ones doing comparative case studies in the first place. Synthetic control places a valuable tool into their hands which enables them to choose counterfactuals - a process that in principle can improve their work insofar as they are interested in evaluating some particular intervention.

Finally, Abadie et al. [2010] argue that it removes subjective researcher bias, but I actually believe this is the most overstated benefit of the method. Through repeated iterations and changes to the matching formula, a person can just as easily introduce subjective choices into the process. Sure, the weights are optimally chosen to minimize some distance function, but through the choice of the covariates themselves, the researcher can in principle select different weights. She just doesn't have a lot of control over it, because ultimately the weights are optimal for a given set of covariates.

Formalization "I'm the real Slim Shady
all you other Slim Shadys are just imitating"
– Eminem

Let Y_{jt} be the outcome of interest for unit j of $J+1$ aggregate units at time t, and treatment group be $j=1$. The synthetic control estimator models the effect of the intervention at time T_0 on the treatment group using a linear combination of optimally chosen units as a synthetic control. For the post-intervention period, the synthetic control estimator measures the causal effect as $Y_{1t}-\sum_{j=2}^{J+1} w_j^* Y_{jt}$ where w_j^* is a vector of optimally chosen weights.

Matching variables, X_1 and X_0, are chosen as predictors of post-intervention outcomes and must be unaffected by the intervention. The weights are chosen so as to minimize the norm, $||X_1 - X_0 W||$ subject to weight constraints. There are two weight constraints. First, let $W=(w_2,\dots,w_{J+1})'$ with $w_j \geq 0$ for $j=2,\dots,J+1$. Second, let $w_2+\dots+w_{J+1}=1$. In words, no unit receives a negative weight, but can receive a zero weight.[139] And the sum of all weights must equal one.

[139] See Doudchenko and Imbens [2016] for recent work relaxing the non-negativity constraint.

As I said, Abadie et al. [2010] consider

$$||X_1 - X_0W|| = \sqrt{(X_1 - X_0W)'V(X_1 - X_0)W}$$

where V is some $(k \times k)$ symmetric and positive semidefinite matrix. Let X_{jm} be the value of the m-th covariates for unit j. Typically, V is diagonal with main diagonal $v_1, \dots, v_k$. Then the synthetic control weights minimize:

$$\sum_{m=1}^{k} v_m \bigg(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \bigg)^2$$

where v_m is a weight that reflects the relative importance that we assign to the m-th variable when we measure the discrepancy between the treated unit and the synthetic control.

The choice of V, as should be seen by now, is important because W^* depends on one's choice of V. The synthetic control $W^*(V)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment. Therefore, the weights $v_1, \dots, v_k$ should reflect the predictive value of the covariates.

Abadie et al. [2010] suggests different choices of V, but ultimately it appears from practice that most people choose V that minimizes the mean squared prediction error:

$$\sum_{t=1}^{T_0} \bigg(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V) Y_{jt} \bigg)^2$$

What about unobserved factors? Comparative case studies are complicated by unmeasured factors affecting the outcome of interest as well as heterogeneity in the effect of observed and unobserved factors. Abadie et al. [2010] note that if the number of pre-intervention periods in the data is "large", then matching on pre-intervention outcomes can allow us to control for the heterogenous responses to multiple unobserved factors. The intuition here is that only units that are alike on unobservables and unobservables would follow a similar trajectory pre-treatment.

California's Proposition 99 Abadie and Gardeazabal [2003] developed the synthetic control estimator so as to evaluate the impact that terrorism had on the Basque region. But Abadie et al. [2010] expounds on the method by using a cigarette tax in California called Proposition 99. The cigarette tax example uses a placebo-based method for inference, which I'm wanting to explain, so let's look more closely at their paper.

In 1988, California passed comprehensive tobacco control legislation called Proposition 99. Proposition 99 increased cigarette taxes by

25 cents a pack, spurred clean-air ordinances throughout the state, funded anti-smoking media campaigns, earmarked tax revenues to health and anti-smoking budgets, and produced more than $100 million a year in anti-tobacco projects. Other states had similar control programs, and they were dropped from their analysis.

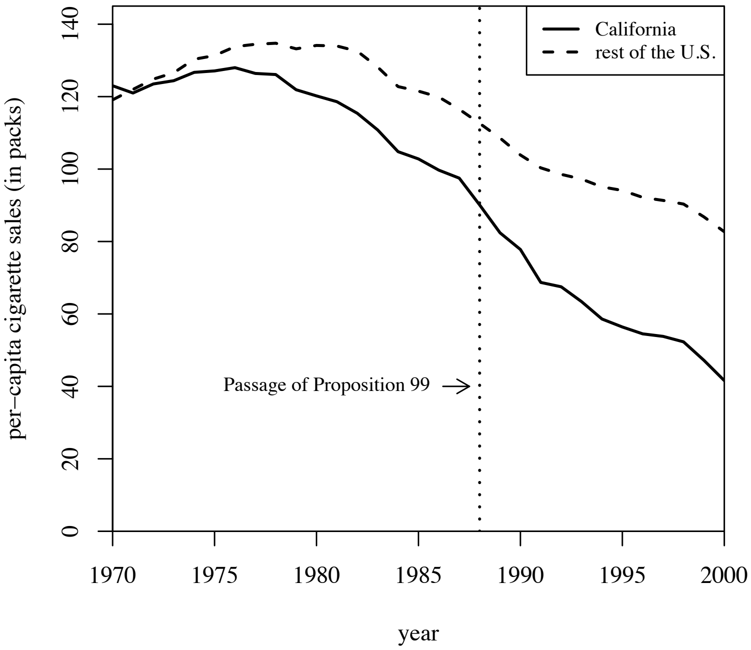

Figure 94: California cigarette sales vs the rest of the country

Figure 94 shows changes in cigarette sales for California and the rest of the United States annually from 1970 to 2000. As can be seen, cigarette sales fell after Proposition 99, but as they were already falling, it's not clear if there was any effect – particularly since they were falling in the rest of the country at the same time.

Using their method, though, they select an optimal set of weights that when applied to the rest of the country produces the figure shown in Figure 95. Notice that pre-treatment, this set of weights produces a nearly identical time path for California as the real California itself, but post-treatment the two series diverge. There appears at first glance to have been an effect of the program on cigarette sales.

The variables they used for their distance minimization are listed in Figure 96. Notice that this analysis produces values for the treatment group and control group that facilitate a simple investigation of balance. This is not a technical test, as there are only one value per variable per treatment category, but it's the best we can do with this method. And it appears that the variables used for matching are similar across the two groups, particularly for the lagged values.

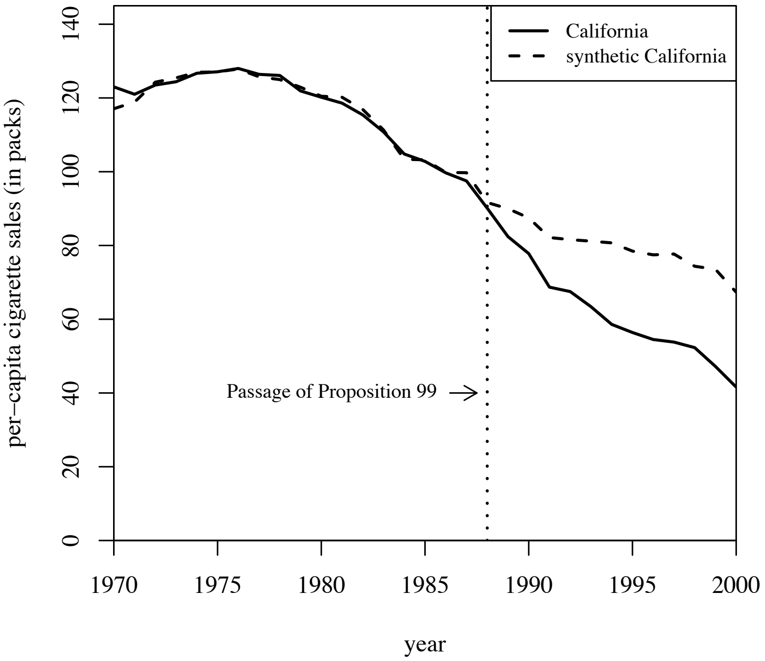

Figure 95: California cigarette sales vs synthetic California

Figure 96: Balance table

	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Like RDD, synthetic control is a picture-intensive estimator. Your estimator is basically a picture of two series which, if there is a causal effect, diverge from another post-treatment, but resemble each other pre-treatment. It is common to therefore see a picture just showing the difference between the two series (Figure 97. But so far, we have only covered estimation. How do we determine whether the observed difference between the two series is a *statistically significant* difference? After all, we only have two observations per year. Maybe the divergence between the two series is nothing more than prediction error, and any model chosen would've done that, even if there was no treatment effect. Abadie et al. [2010] suggest that we use an old fashioned method to construct exact p-values based on Fisher [1935]. This is done through "randomization" of the treatment to each unit, re-estimating the model, and calculating a set of root mean squared prediction error (RMSPE) values for the pre- and post-treatment period.[140] We proceed as follows:

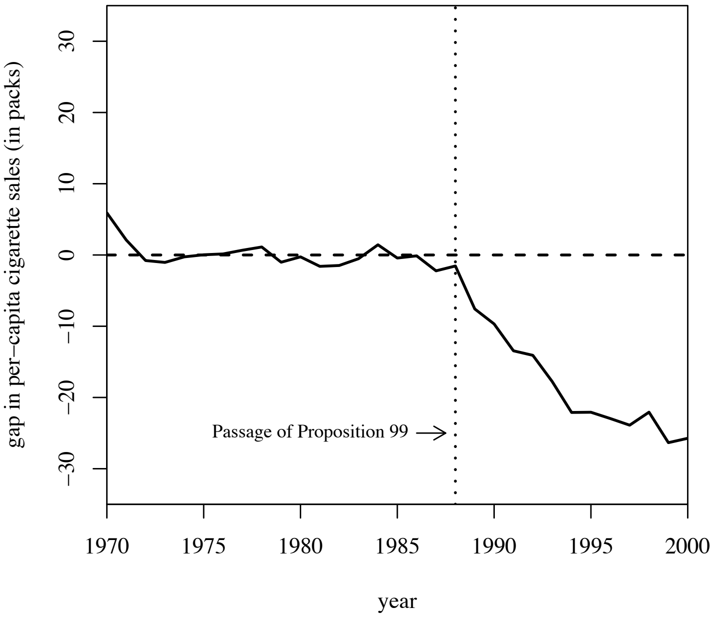

Figure 97: California cigarette sales vs synthetic California

1. Iteratively apply the synthetic control method to each country/state in the donor pool and obtain a distribution of placebo effects

2. Calculate the RMSPE for each placebo for the pre-treatment period:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+t}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

[140] What we will do is simply reassign the treatment to each unit, putting California back into the donor pool each time, estimate the model for that "placebo", and recording information from each iteration.

3. Calculate the RMSPE for each placebo for the post-treatment period (similar equation but for the post-treatment period)

4. Compute the ratio of the post-to-pre-treatment RMSPE

5. Sort this ratio in descending order from greatest to highest.

6. Calculate the treatment unit's ratio in the distribution as $p = \frac{RANK}{TOTAL}$

In other words, what we want to know is whether California's treatment effect is extreme, which is a relative concept compared to the donor pool's own placebo ratios.

There's several different ways to represent this. The first is to overlay California with all the placebos using Stata `twoway` command, which I'll show later. Figure 98 shows what this looks like. And I think you'll agree, it tells a nice story. Clearly, California is in the tails of some distribution of treatment effects. Abadie et al. [2010] recommend iteratively dropping the states whose pre-treatment RMSPE is considerably different than California's because as you can see, they're kind of blowing up the scale and making it hard to see what's going on. They do this in several steps, but I'll just skip to the last step (Figure 99). In this, they've dropped any state unit from the graph whose pre-treatment RMSPE is more than two times that of California's. This therefore limits the picture to just units whose

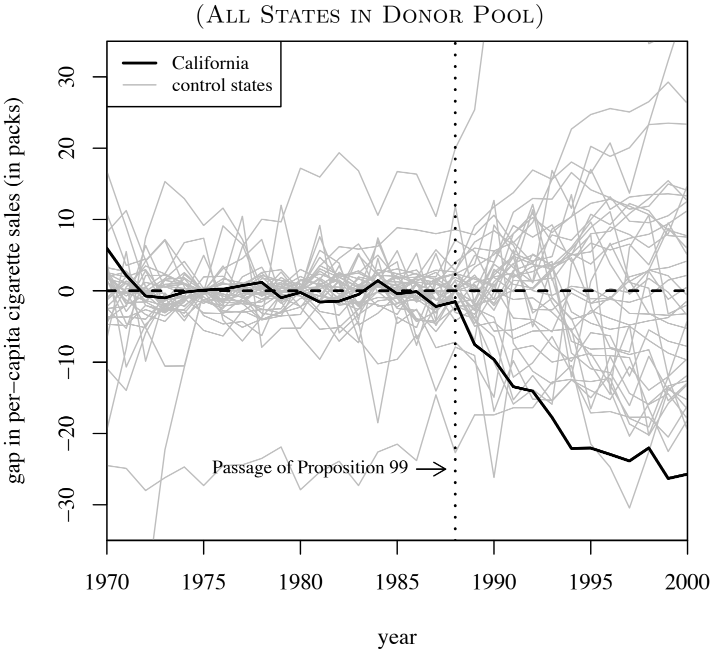

Figure 98: Placebo distribution

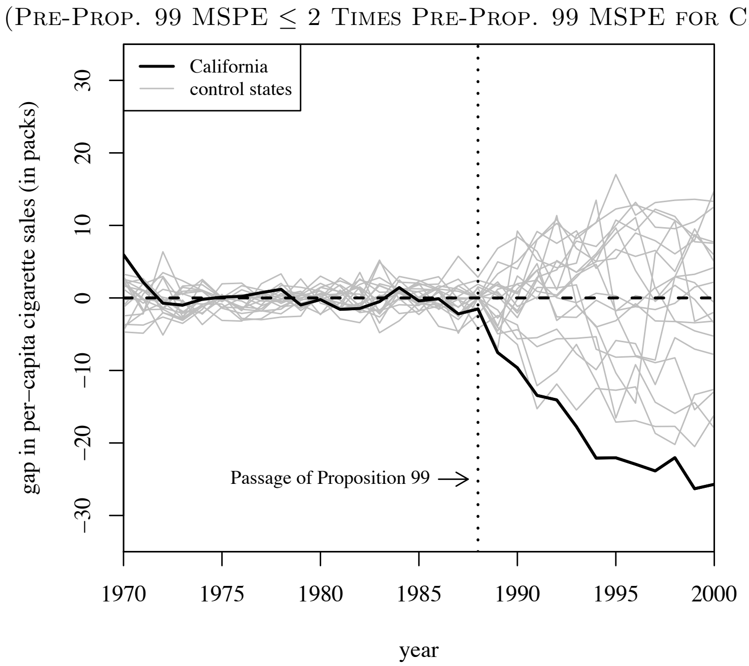

Figure 99: Placebo distribution

model fit, pre-treatment, was pretty good, like California's. But, ultimately, inference is based on those exact p-values. So the way we do this is we simply create a histogram of the ratios, and more or less mark the treatment group in the distribution so that the reader can see the exact p-value associated with the model. I produce that here in Figure 100. As can be seen, California is ranked 1st out of 38 state units.[141] This gives an exact p-value of 0.026, which is less than the conventional 5% most journals want to (arbitrarily) see for statistical significance.

Falsifications In Abadie et al. [2015], the authors studied the effect of the reunification of Germany on GDP. One of the contributions this paper makes, though, is a recommendation for how to test the validity of the estimator through a falsification exercise. To illustrate this, let's walk through their basic findings. In Figure 101, the authors illustrate their main question by showing the changing trend lines for West Germany and the rest of their OECD sample.

As we saw with cigarette smoking, it's difficult to make a statement about the effect of reunification given West Germany is dissimilar from the other countries on average before reunification.

In Figure 101 and Figure 103, we see their main results. The authors then implement the placebo-based inference to calculate exact $p-$-values and find that the estimated treatment effect from reunifica-

[141] Recall, they dropped several states who had similar legislation passed over this time period.

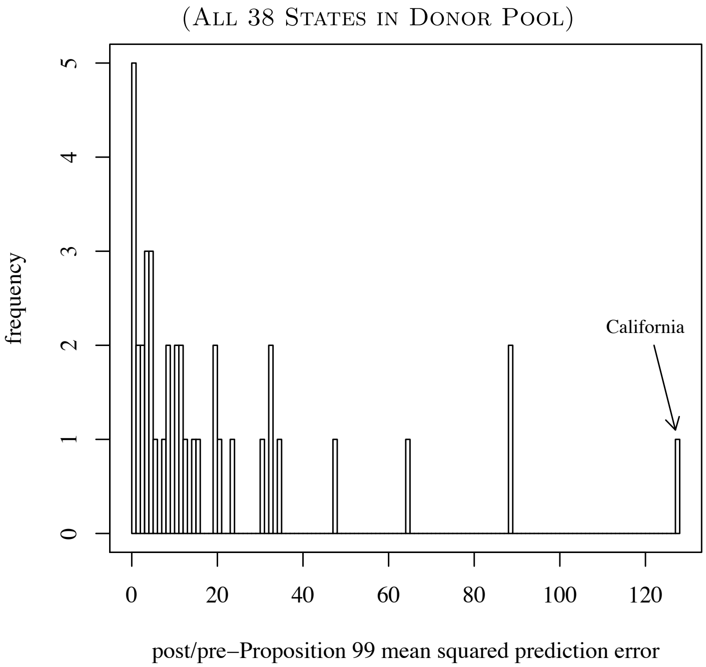

Figure 100: Placebo distribution

Figure 1: Trends in Per-Capita GDP: West Germany vs. Rest of OECD Sample

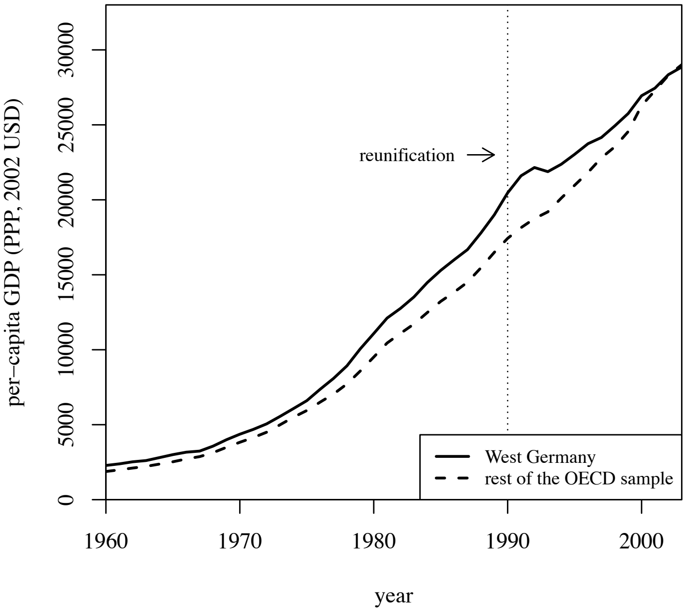

Figure 101: West Germany GDP vs. Other Countries

Figure 2: Trends in Per-Capita GDP: West Germany vs. Synthetic West Germany

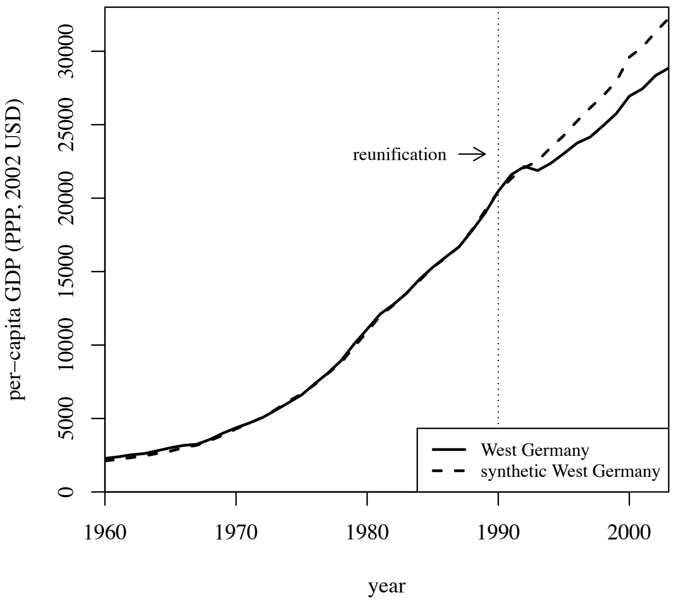

Figure 102: Synthetic control graph: West Germany vs Synthetic West Germany

Figure 3: Per-Capita GDP Gap Between West Germany and Synthetic West Germany

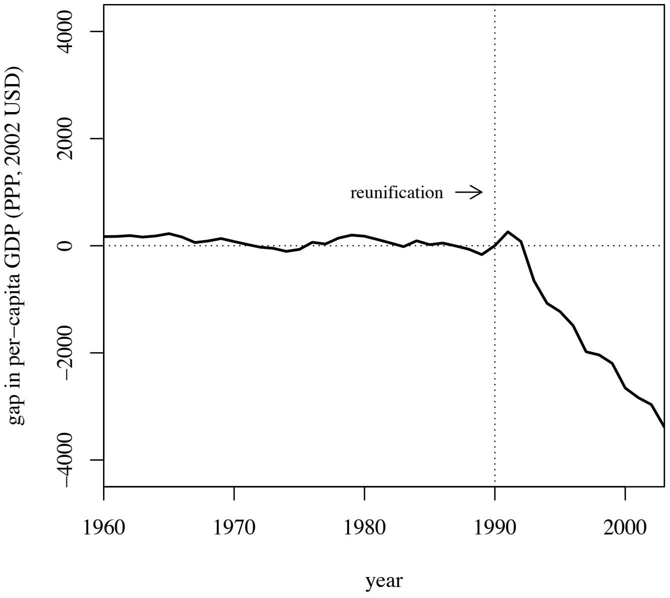

Figure 103: Synthetic control graph: Differences between West Germany and Synthetic West Germany

tion is statistically significant.

The placebo-based inference suggests even further robustness checks, though. The authors specifically recommend rewinding time from the date of the treatment itself and estimating their model on an earlier (placebo) date. There should be no effect when they do this; if there is, then it calls into question the research design. The authors do this in Figure 104. Notice that when they run their model on the placebo date of 1975, they ultimately find no effect. This suggests that their model has good in and out of sample predictive properties. Hence since the model does such a good job of predicting GDP per capita, the fact that it fails to anticipate the change in the year of reunification suggests that the model was picking up a causal effect.

Figure 4: Placebo Reunification 1975 - Trends in Per-Capita GDP: West Germany vs. Synthetic West Germany

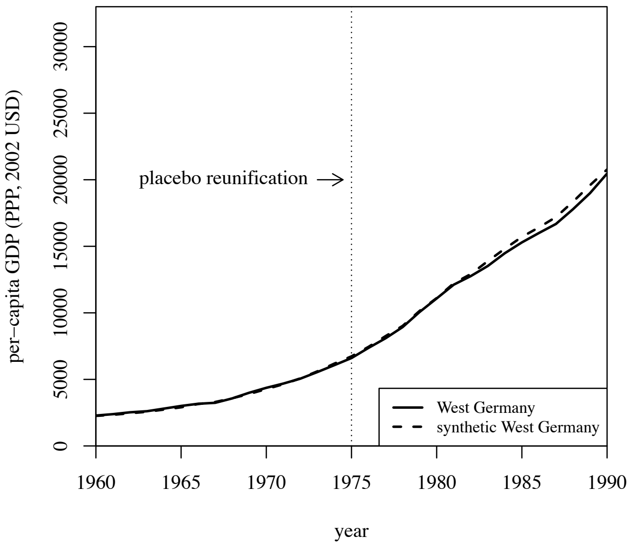

Figure 104: Synthetic control graph: Placebo Date

We include this second paper primarily to illustrate that synthetic control methods are increasingly expected to pursue numerous falsification exercises in addition to simply estimating the causal effect itself. In this sense, researchers have pushed others to hold it to the same level of scrutiny and skepticism as they have with other methodologies such as RDD and IV. Authors using synthetic control must do more than merely run the `synth` command when doing comparative case studies. They must find the exact p-values through placebo-based inference, check for the quality of the pre-treatment fit, investigate the balance of the covariates used for matching, and check for the validity of the model through placebo estimation (e.g., rolling back the treatment date).

Stata exercise: Prison construction and Black male incarceration

The project that you'll be replicating here is a project I have been working on with several coauthors over the last few years.[142] Here's the backdrop.

In 1980, Texas Department of Corrections lost a major civil action lawsuit. The lawsuit was called *Ruiz v. Estelle*; Ruiz was the prisoner who brought the case, and Estelle was the warden. The case argued that TDC was engaging in unconstitutional practices related to overcrowding and other prison conditions. Surprisingly, Texas lost the case, and as a result, Texas was forced to enter into a series of settlements. To amend the issue of overcrowding, the courts placed constraints on the number of housing inmates that could be placed in cells. To ensure compliance, TDC was put under court supervision until 2003.

Given these constraints, the construction of new prisons was the only way that Texas could adequately meet demand without letting prisoners go, and since the building of new prisons was erratic, the only other option was increasing the state's parole rate. That is precisely what happened; following *Ruiz v. Estelle*, Texas used paroles more intensively to handle the increased arrest and imprisonment flows since they did not have the operational capacity to handle that flow otherwise.

But, then the state began building prisons which started somewhat in the late 1980s under Governor Bill Clements. However, the prison construction under Clements was relatively modest. Not so in 1993 when Governor Ann Richards embarked on a major prison construction drive. Under Richards, state legislators approved a billion dollar prison construction project which doubled the state's operational capacity within 3 years. This can be seen in Figure 105. As can be seen, Clements build out was relatively modest both as a percentage change and in levels. But Richards' investments in operational capacity was gigantic – the number of beds grew over 30% for three years causing the number of beds to more than double in a short period of time.

What was the effect of building so many prisons? Just because prison capacity expands doesn't mean incarceration rates will grow. But because the state was intensively using paroles to handle the flow, that's precisely what did happen. Because our analysis in a moment will focus on African-American male imprisonment, I will show the

[142] You can find one example of an unpublished manuscript here coauthored with Sam Kang: http://scunning.com/prison_booms_and_drugs_20.pdf