Inclusion of the treated unit in the synthetic control pool: Two case studies

Akmarzhan Abylay, Frederik Hardervig and Natalia Velasco

CS112 Spring 2020

To: Authors and peer reviewers of papers who use synthetic control methods

From: Akmarzhan Abylay, Frederik Hardervig, and Natalia Velasco

Date: April 24th, 2020

Re: Analyzing the effect of including the treated unit into the synthetic control donor pool

Executive Summary

This decision memo proposes including the treatment effect into the synthetic control donor pool when performing rigorous hypothesis testing (e.g., sharp null) and excluding/including the unit of interest based on the preferences in other papers not based on the assumption of no treatment effect. Due to no standardized procedure and a present dilemma of including/excluding the treatment unit, as well as a lack of papers examining its effects on the significance of the results, we decided to perform a variety of tests and identify the actual significance of adding the unit of interest into the synthetic control pool. Our proposal will allow researchers to save resources on exploring the effects of the treatment unit inclusion, as well as strengthen the reliability of prospective treatment effect tests, which could add to the field of causal inferences in future.

Background

The synthetic control method is useful when evaluating the effect of intervention on specific case studies, where perfect control units with equivalent characteristics are implausible to exist. While a rigorous synthetic control approach can allow for the calculation of treatment effect, it is important to calculate the significance of such estimates to ensure our conclusions are appropriate.

To do this, placebo tests are run by iteratively applying the synthetic control method to each unit in the donor pool, and computing the estimated effect. This process yields a distribution of estimated gaps for the donor units, and if these gaps are similar in magnitude to the one originally estimated for the treated unit, then we have reason to believe that the treatment effect is not extreme enough to be considered significant.

In this procedure, the composition of the "donor pool" used to calculate placebo tests with other units can vary slightly between applications. One approach is to adhere to the assumption that the null hypothesis is true, implying that the intervention had no effect on the outcome of the unit of interest, and therefore include the treated unit in the placebo tests. This is exemplified in Abadie, Diamond, and Hainmueller (2012), who study the effects of Proposition 99 (adding a tax on tobacco) on cigarette sales in California[1]. When creating the placebo tests, these authors included California in the donor pool. This is consistent with their assumption of no treatment effect.

A different approach is the one exemplified by Abadie and Gardeazabal (2003) in their study of the economic effects of the terrorist activity that emerged in the Basque Region in the late 1960's[2]. The paper estimates an effect of 10% decrease in per capita GPD due to terrorism, and then conducts a placebo test to "assess whether the gap observed for the Basque Country may have been created by factors other than terrorism" (Abadie & Gardeazabal, 2003). The authors use Cataluña as a comparison unit, but exclude the treated unit (Basque Country) in the construction of "synthetic Cataluña." This design choice implies that the authors did not have a sharp null hypothesis, and could have considered some effects in the Basque country that could confound synthetic Cataluña. The sample size, that was merely 18 units, limits the possibility of reaching significant p-values..

In this paper, we explore the implications of both alternatives on these analyses, and what we can learn for future, similar case studies.

Replication

In the paper about "Estimating the Effect of California's Tobacco Control Program" (Abadie, Diamond, & Hainmueller, 2012), the authors created a plot of the gap between per capita cigarette consumption in California compared to "synthetics California", including the periods before and after the passage of Proposition 99 (Appendix B). They estimated a large impact of this piece of legislation, with the cigarette consumption decreasing by almost 25% after its passage. They proceed to then plot the same gaps for the other 38 control states considered, making it easy for readers to visually compare these estimated effect magnitudes. With the same data, we were able to replicate their findings and produce an identical plot of *gap of per capita cigarette sales, in packs* vs. *year.*

[1] This paper will also be referred to as "California paper" throughout this text, for simplicity and to avoid confusion between two papers co-authored by Abadie.

[2] This paper will also be referred to as "Basque Country paper" throughout this text, for the same reasoning explained before.

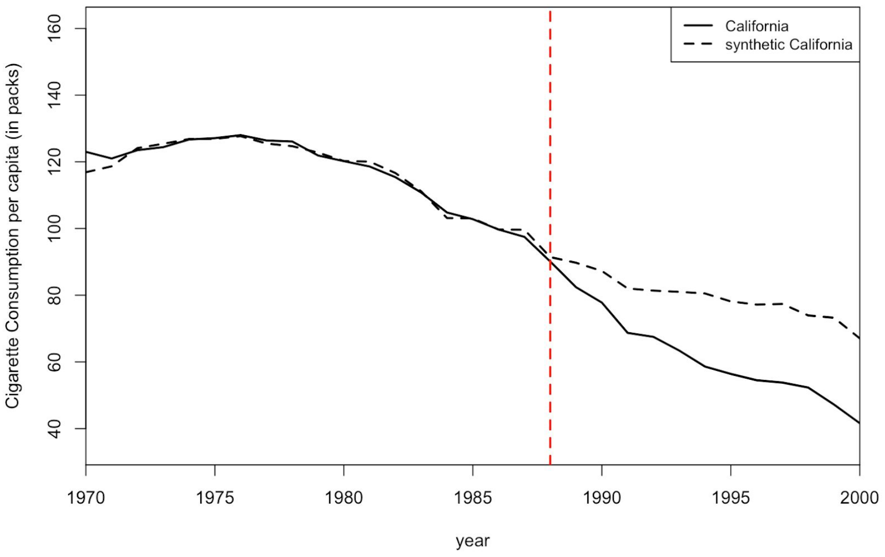


Figure 1. *Replication of "Trends in per-capita cigarette sales: California vs. synthetic California" (Figure 2) in Abadie, Diamond, & Hainmueller (2012). The line for California is bolded and matches the line for synthetic California until after the passage of Proposition 99. See Appendix B for original figures.*

To evaluate the significance in their estimated treatment effect, the authors conducted a series of placebo tests on other states who did not implement anti-tobacco legislation. The authors refine the scope of the placebo test by excluding states with significantly larger smoking rates than California (pre-intervention), but include California as a donor for other states' tests.

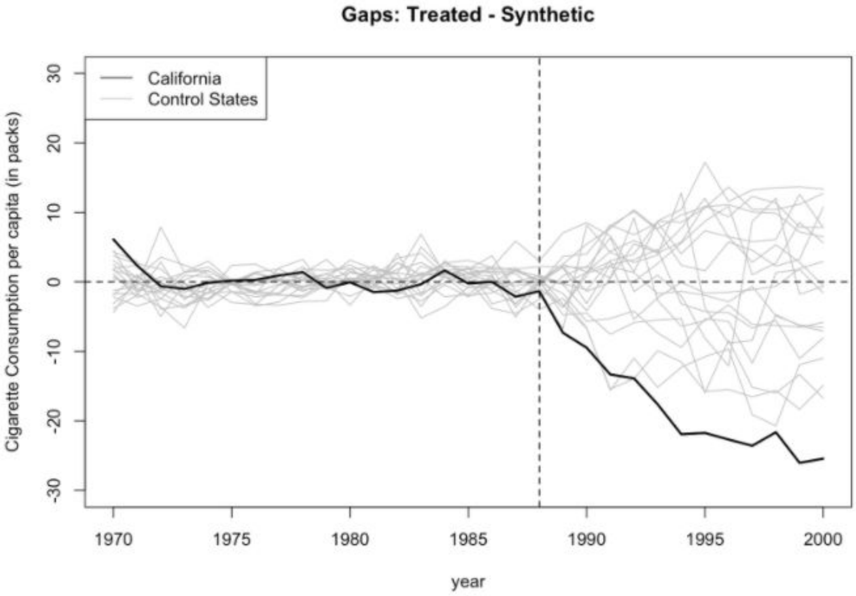


Figure 2. *Replication of "Per-capita cigarette sales gaps in California and placebo gaps in 19 control states" (Figure 7) in the California paper. The line for California is bolded, and 19 control*

states are depicted (the states with pre-Proposition 99 MSPE two or more times higher than California's are discarded). See Appendix B for original figures.

A similar procedure was followed to replicate "A Case Study of the Basque Country" (Abadie & Gardeazabal, 2003). The authors provided a figure in which they showed the real-per capita GDP for both the Basque country and the Synthetic Basque.

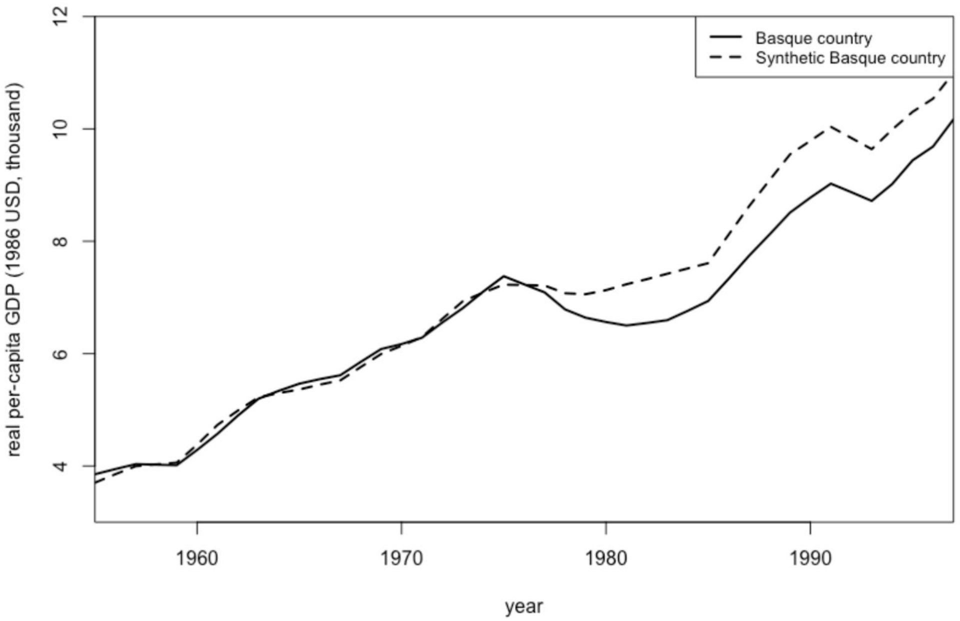


***Figure 3**. Replication of "Per Capita GDP for the Basque Country"(Figure 1) from Abadie & Gardeazabal (2003). As we can see both the Basque Country and its synthetic control lines observed a similar pattern until 1975. This is the year labeled by the authors as the time when "ETA's terrorist activity becomes a large-scale phenomenon" (Abadie & Gardeazabal, 2003).* See Appendix C for original figures.

The authors find an effect size of up to 12% reduction in GDP per capital due to terrorism, represented by the gap between the real and synthetic lines in the plot. Then, a placebo test is conducted to asses if the effect might have been due to chance, or other variables that are not terrorism. As discussed previously, this placebo test uses Cataluña, and a synthetic Cataluña created by other spanish regions and excluding the Basque Country.

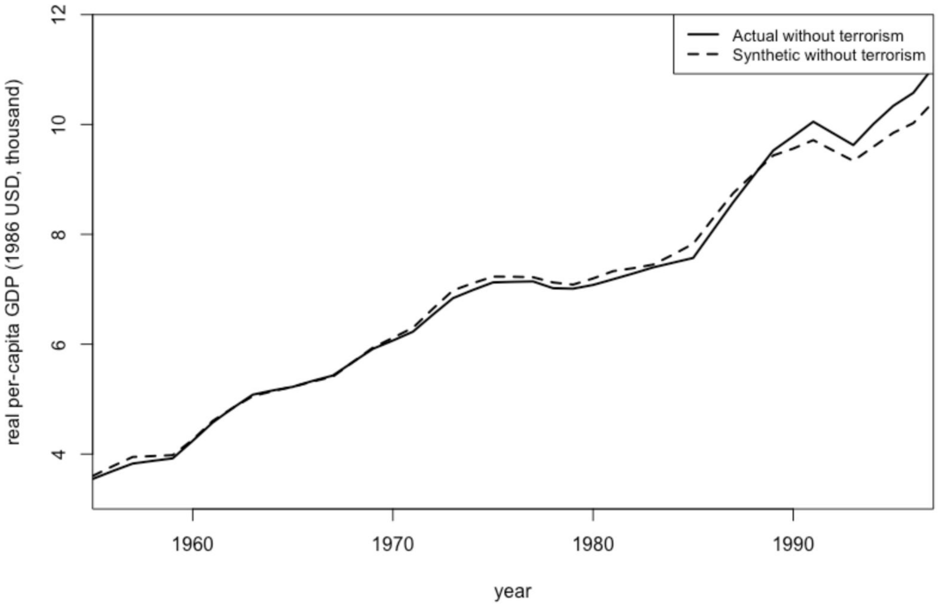


Figure 4*. Replication of "*<u>*A Placebo Study, per capita GDP for Catalonia*</u>*" (Figure 4) from the Basque Country paper. In this case, Cataluña and its synthetic control lines observed a similar pattern throughout the period in question (except in 1992, when the Olympics in Barcelona led to economic expansion). Given that Cataluña did not experience widespread terrorism, this placebo test provides evidence that terrorist activity might have plausibly caused a lag in the Basque Country economy. See Appendix C for original figures.*

An extension of these replications would involve exploring how these placebo test would change if the alternative design choices were made; ie. include the treatment unit in the donor pool when it was originally excluded, and vice versa.

Problem Statement

Our primary motivation behind this study is to explore the significance of including the treatment unit in the donor pool. E.g. in the California paper, the author's are exploring a sharp null hypothesis, thus assuming that California is not significantly different from the other units. However, after their analysis they concluded that that it is significantly different, and thus one could argue that it should not be included in the donor pools for the placebo tests since it might negatively influence the accuracy of the placebo synthetic control post-treatment, thus leading to less certainty in the quality of the synthetic control built for the treatment unit. On the other hand, the Basque paper assumes from the beginning that terrorism has had an effect in Basque country, and thus doesn't use it in its placebo tests. This limits the donor pool to just 18 units, which means that even a single more synthetic control might have a significant impact compared to studies with many more donor units. In the worst case, we suspect that there may be a significant difference in the accuracy of our placebo tests depending on whether our treated unit is included in the donor pool for the synthetic controls. This would raise a set of questions that researchers must ask themselves before performing placebo tests:

- Does our question lend itself to including the treatment unit in the donor pool?
- What may be gained from including the treated unit?
- Might it invalidate the synthetic controls?
- And finally: Should the treated unit be included in the donor pool for the placebo tests?

These aren't necessarily trivial questions, and thus we will now seek to set up some heuristics and tests in order to help decide whether to include the treated unit or not.

Based on the preliminary knowledge about synthetic controls, we assume that including the treatment unit into the donor pool would introduce additional variance, which could potentially skew results. In order to check this hypothesis, we decided to analyze some papers which included/excluded the treatment unit and explore whether performing an opposite procedure would yield any significant difference to the results or would introduce a significant variance as we have assumed in the beginning.

We replicated two papers as mentioned in the previous sections and tried to analyze the results of our manipulations. In the Basque country paper the researchers had already assumed that a treatment effect had taken place in Basque country, and thus excluded it from the placebo donor pool. However, we will explore what might have happened if we did include Basque country in the donor pool. Will it introduce post-treatment variance in the placebo synthetic controls or will the variance reduce as we can better fit with more donors?

To answer these questions, we will evaluate the effect of including Basque country in the donor pool on those placebo synthetic controls that use it if available.

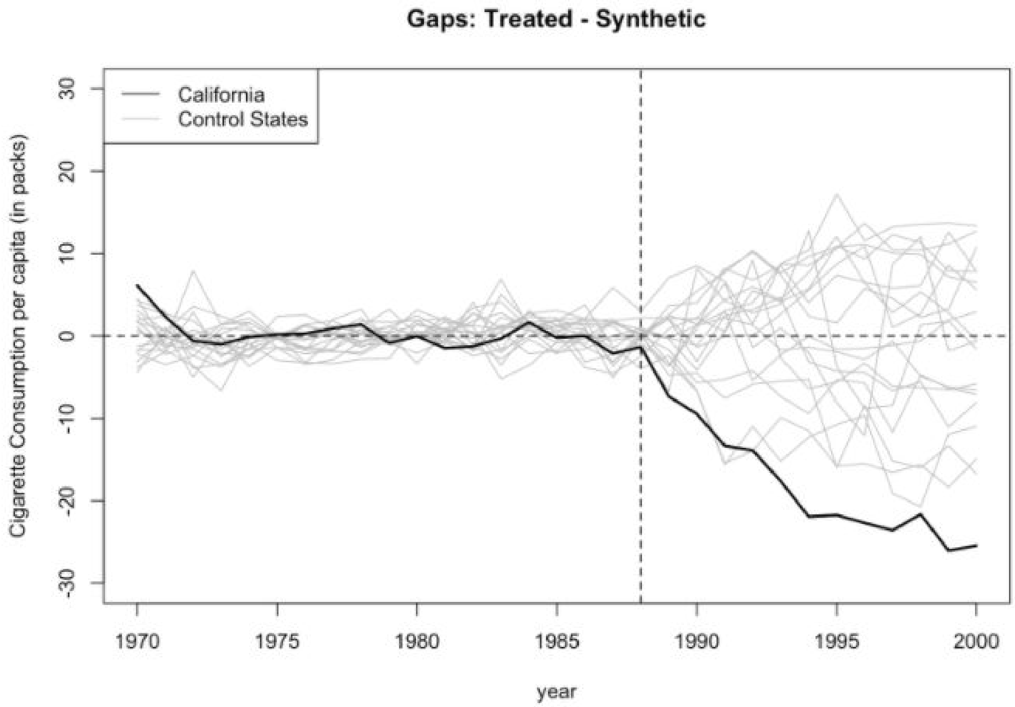


Figure 5. *The figure above shows a replication of Figure 7 in the California paper. It is a gap plot, showing the discrepancies between control states and their placebo synthetic control in gray, and discrepancies between California and its synthetic control in black.*

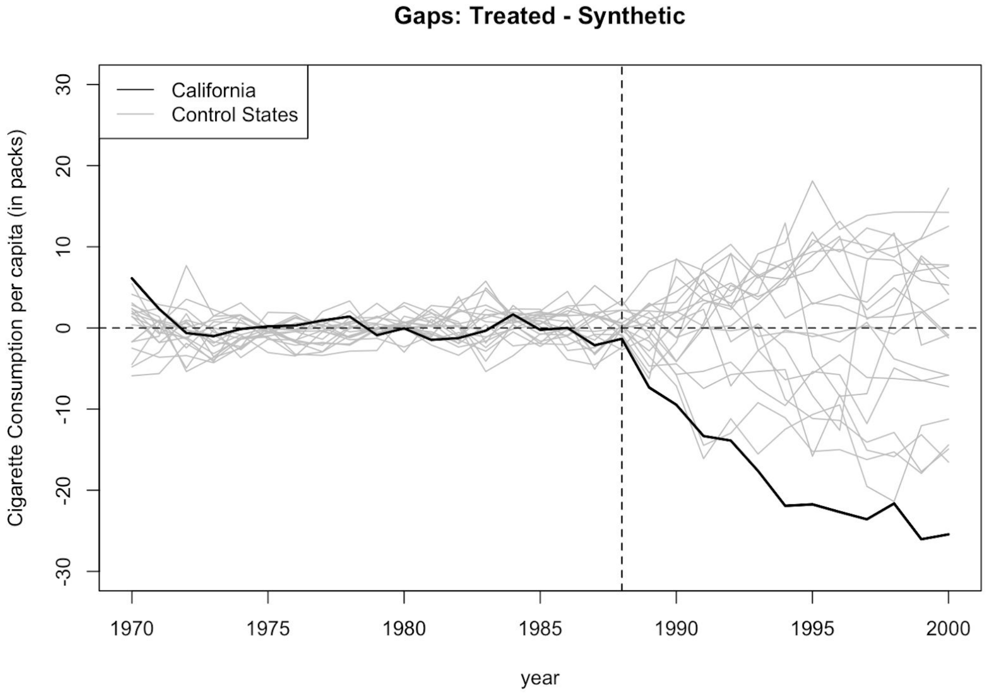


Figure 6. *This figure shows the same as our replication figure, except the synthetic controls no longer have California in their donor pool. Many of the placebos are unchanged as they weren't*

using California in the first place, so we will now remove all placebos using less than 1% California and compare.

Qualitatively, we can see minor differences between figure 5 and 6, so we will now calculate the mean squared summed error, mean summed error, and standard deviation in the post treatment period for both figures. The reason we are summing the errors for all years in the post-treatment period is to get a more holistic picture of the influence of including/excluding CA during the post treatment period, rather than just looking at the end-result. The table below shows that we found no significant difference in the mean squared summed errors. While insignificant for the overall sample of units, not all units wish to use CA for their synthetic control even when available, so the increase in error when including CA vs. excluding it, may stem from a few units where it has significant impact. Thus we will now focus on the placebos using more than 1% CA in their synthetic control, versus those who used less.

IMPACT OF INCLUDING/EXCLUDING CALIFORNIA

	Mean Summed Error	MSSE	SD
WITH ACCESS TO CA	115.32	46501.56	169203.8
W.O. ACCESS TO CA	107.04	41721	164082.4
DIFFERENCE:	8.28	4780.56	5121.4
SIGNIFICANCE:		**p=0.9008**	

Figure 7. *The impact of excluding or excluding California in the placebo tests on three different metrics (n=38). For further detail about the significance, see Appendix D.*

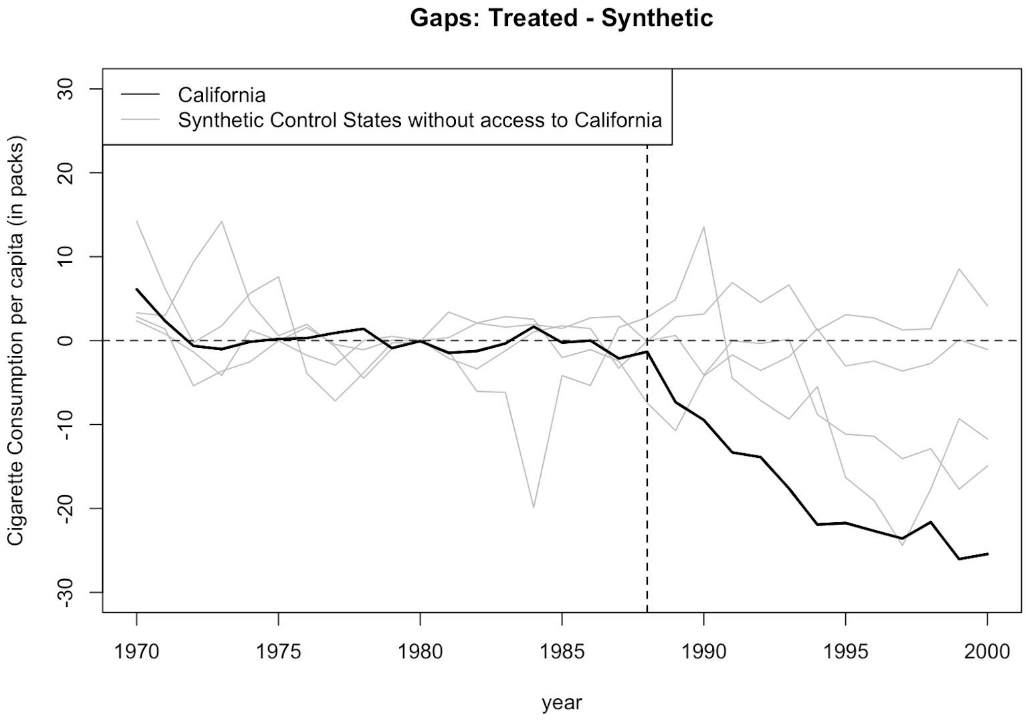


Figure 8. *This figure shows the placebos of the states that would like to use more than 1% of California in their synthetic control if they had access to it.*

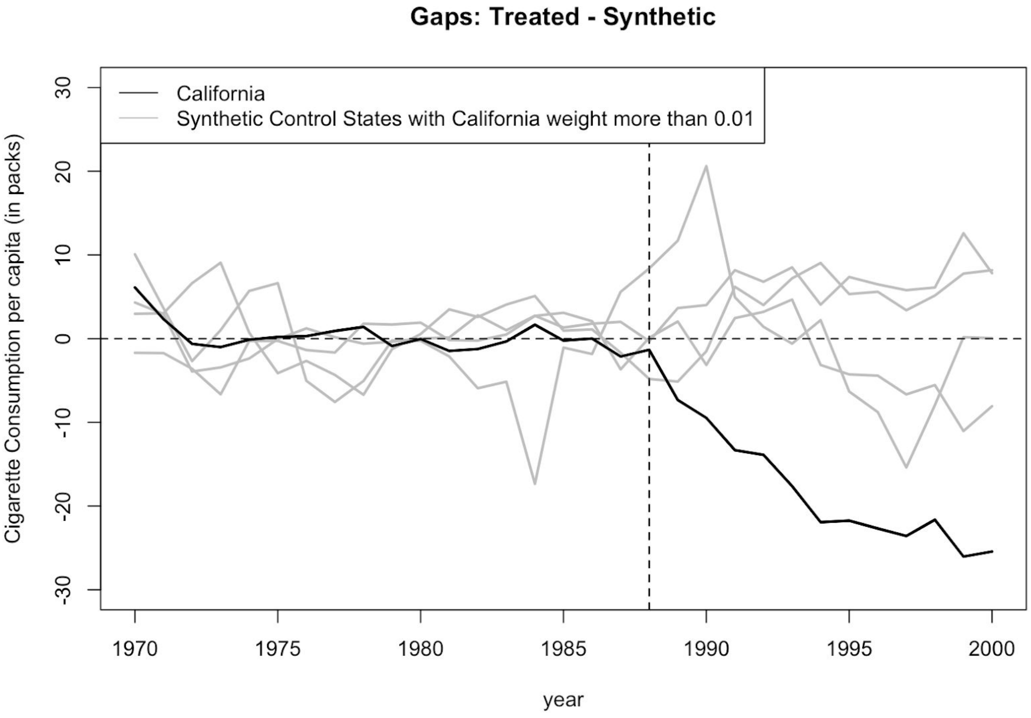
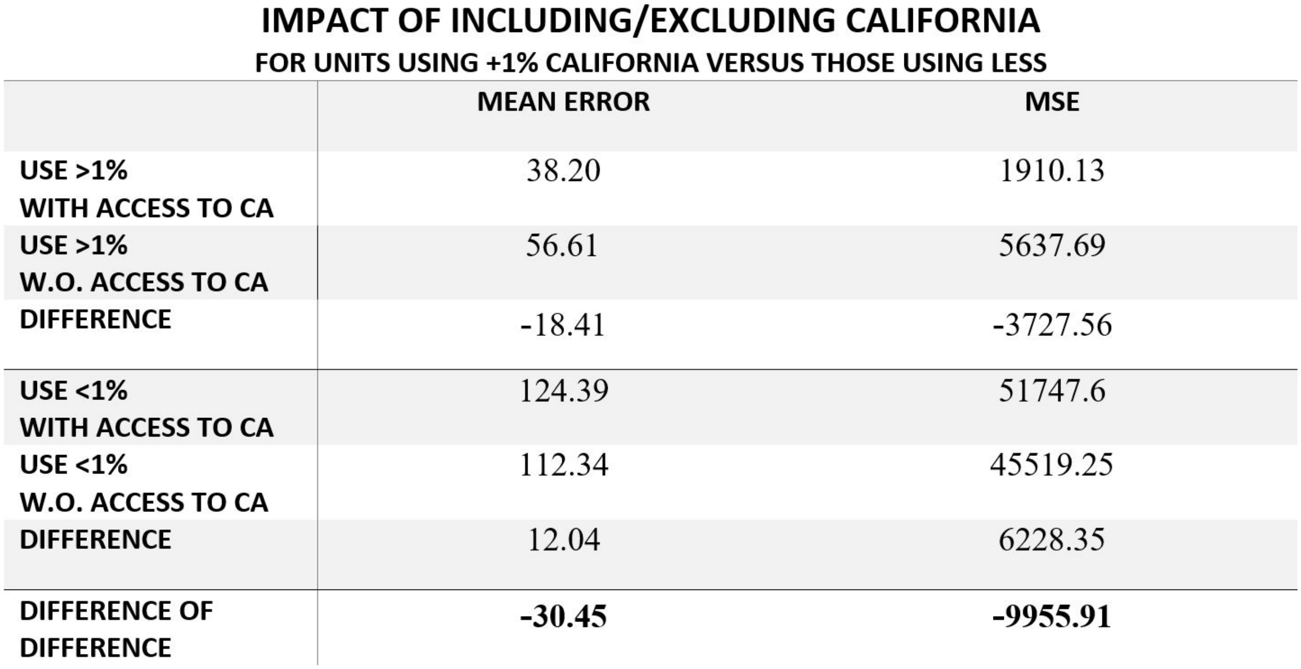


Figure 9. *Here we have all the placebos where more than 1% of the synthetic control was made up of California. While there are clear differences, we will choose to compare them quantitatively rather than qualitatively.*



IMPACT OF INCLUDING/EXCLUDING CALIFORNIA

FOR UNITS USING +1% CALIFORNIA VERSUS THOSE USING LESS

	MEAN ERROR	MSE
USE >1% WITH ACCESS TO CA	38.20	1910.13
USE >1% W.O. ACCESS TO CA	56.61	5637.69
DIFFERENCE	-18.41	-3727.56
USE <1% WITH ACCESS TO CA	124.39	51747.6
USE <1% W.O. ACCESS TO CA	112.34	45519.25
DIFFERENCE	12.04	6228.35
DIFFERENCE OF DIFFERENCE	**-30.45**	**-9955.91**

Figure 10. *The impact of excluding or excluding California in the placebo tests when split into groups wishing to use more (n=4) or less (n=34) than 1% of California in their synthetic control, given availability of California. We see that our initial suspicion that the overall increase to variance stemmed from a few units being highly skewed by California is incorrect, as those using California saw a decreased error in the post-treatment period.*

The results in the table above seem somewhat counterintuitive. The units using a little California in their placebo synthetic controls see an increase to their variance, while those using more see decreased variance in the post-treatment period. To explore this further, we began exploring the change in weights of the units using more than 1% California. We noticed a common pattern, that apart from using California when available, the weight of unit 36 and similar units was increased drastically, having a weight of approximately 15%-20% across the units. For reference, unit 36 has one of the largest consumption of cigarettes in the dataset with a post-treatment average of 107.8 packs per capita, which we suspect might even out the decrease in consumption from including more of California with its post-treatment mean of 60.35 package per capita. However, we did not explore this further, but instead note that the inclusion of California itself did not introduce/decrease the variance significantly, compared to the impact the shifting of weights had on other units. Along with our insignificant P-value, this is further evidence that there isn't a significant difference to including/excluding the treatment unit.

To further explore the significance of inclusion/exclusion of the treatment unit, we also looked at the Basque Country paper. And found the following results in post treatment MSE and post treatment mean error.

Impact of Including/Excluding Basque Country

	MSE	SD	AE	AE SD
With access to Basque	0.08633435	0.2178838	0.1507217	0.2604966
W.o. Access to Basque	0.09427279	0.3083031	0.1362377	0.2841822
Difference:	- 0.00793844	-0.0904193	0.0144840	-0.0236856
Significance:	**p=0.9335**		**p=0.8815**	

Again, we see that the difference in both the mean squared error and mean error is insignificant, just like in the California paper, so we have no reason to conclude anything else, than that there is no significant difference between including and excluding the treatment unit in the donor pool, although per principle, we would always include it when do a sharp null hypothesis test.

Proposal

We propose to decide on the inclusion of the treatment variable depending on the type of the hypothesis and study's purpose. In other words, based on the results of our analysis in the previous section, we identified that there was no significant difference between the variances of the gaps between the pseudo-treated unit and its synthetic control in the donor pools, which included and excluded the unit of treatment (i.e., Basque country and California in their respective papers). That is why, we propose to include the unit of interest into the donor pool when performing rigorous hypothesis tests, with a null hypothesis of no treatment effect. However, it is completely up to authors of papers if they want to include the treatment unit in other studies, which don't base their analysis on the assumption of no treatment effect.

When making causal inference conclusions, different scientists have different initial assumptions towards the effect of the treatment on the outcome. On one hand, some assume that there is no treatment effect and set a sharp null hypothesis as representing no difference in the outcomes of the treated and the control groups as seen from the California paper. This way it is easier to prove whether the treatment effect is large enough to be considered significant through collecting enough evidence to reject the null hypothesis and accept the alternative one (i.e., that there is treatment effect). As the main premise is that there is no difference between the two groups, ideally, there should be no effect of including the treatment unit into the donor pool for the synthetic control choice. As explained in the sections above, the synthetic control units are constructed as weighted combinations of the units in the donor pool based on the characteristics of the treatment unit before the treatment. Thus, if there is no treatment effect, the actual treatment unit and the other control units should be somewhat similar and even if we include the actual treatment unit into the donor pool and some synthetic controls end up using it, their outcomes after the treatment should not be different from the pseudo-treatment[3] unit that it was matched to.

However, as found in the problem section, the variance that it adds to the observations is not significant enough to cause any difference into the results of the studies. Some differences may arise if there are some synthetic controls that are fully assembled from the treatment unit and the difference between the pseudo-treatment and the synthetic control is the same as the difference between the treatment unit and its corresponding synthetic control (i.e., the actual treatment effect), which should be very rare. Depending on the dataset, the inclusion of the treatment unit can either improve or worsen the variation on the control units that use a considerable (i.e., more than 1%) amount of the treatment unit for the construction of their synthetic control due to the reassignment of weights (e.g., improved in California paper and worsened in Basque country paper). Thus, when following the hypothesis testing framework, it is better to include the treatment unit into the donor pool for the consistency of assumptions.

On the other hand, depending on the purpose and the scope of the study, some people don't choose the hypothesis testing framework, or use it with an assumption other than the absence of the treatment effect. In these cases, it is completely up to the authors whether to include the treatment

[3] This notation is used to refer to the control units to which the synthetic control is found - in the code, they are used as the temporary treatment unit, but they are not the actual treatment unit.

unit or discard it as there is no clear assumption of no treatment effect and there wouldn't be an inconsistency, as well as the inclusion/exclusion doesn't really change the resulting findings.

It is important to make these distinctions, as the exclusion of the treatment unit in specific hypothesis tests could be contrary to the main assumptions. However, for papers of other scopes, the inclusion/exclusion of the unit of interest from the donor pool for the synthetic control wouldn't make any difference to the final results.

Conclusion

In this paper, we analyzed the dilemma of including/excluding the treatment unit from the donor pool for the synthetic controls when exploring the treatment effect or the significance of the findings through the placebo tests. We have replicated the placebo test methods of two publications, one that includes treated units and one that excludes them. This paper reveals that it is vital to have the purpose of the study in mind and remain consistent with the design choices: in other words, if the purpose is to conduct a hypothesis test with an assumption of no treatment effect, then it is sensible to include the treated unit to avoid inconsistencies in the assumptions. On the other hand, if the study doesn't include any assumptions connected to no treatment effect, it should make no significant difference between the two choices, which means that it is up to the authors to include it or exclude it.

Still, authors need to be aware that results can change based on how they decide to design their placebo tests. We found an **insignificant** difference in the results for both papers when we tested the effect of the treatment unit inclusion on the variance of the observations. In the case of the California paper, we got a p-value of 0.9008 from the difference in means test, while the same test for Basque paper had a slightly better but still largely insignificant p-value of 0.8815.

We have come to a conclusion that future papers need to show some consideration for this nuance. However, there is still room for exploration and we are open to consideration of new evidence.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and synthetic control methods. American Journal of Political Science, 59(2), 495-510. Retrieved from https://www.researchgate.net/publication/228322749_Comparative_Politics_and_the_Synthetic_Control_Method/download.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. American Economic Review, 113-132. Retrieved from https://economics.mit.edu/files/11870.

MedCalc. (2020). Comparison of means calculator. Retrieved from https://www.medcalc.org/calc/comparison_of_means.php.

APPENDIX

You can find the data here. You can find all code in this Github gist and it is also uploaded as a secondary file.

Appendix A - HC Applications

- **#controlgroups:** The synthetic control method essentially aims to create the best possible "fake" control group for a studied unit, and we accurately identified how the authors go about using these synthetic controls in their respective empirical studies. We evaluated their design choices when constructing these control groups and provided a detailed critique of how, if at all, changing the composition of these control groups to include/exclude the treated unit could change the conclusions.

- **#significance**: We replicated the tests of statistical significance found in both papers, calculating p-values using placebo tests, explaining how the assumptions about the null hypothesis could change the way these tests were conducted (including vs excluding the treated variable). We also conducted the difference in means significance tests to identify whether the difference between the variances of the observations with an included/excluded treatment unit was significant enough to claim that it would severely impact the observations introducing additional variance.

Appendix B - Original figures from the California Paper

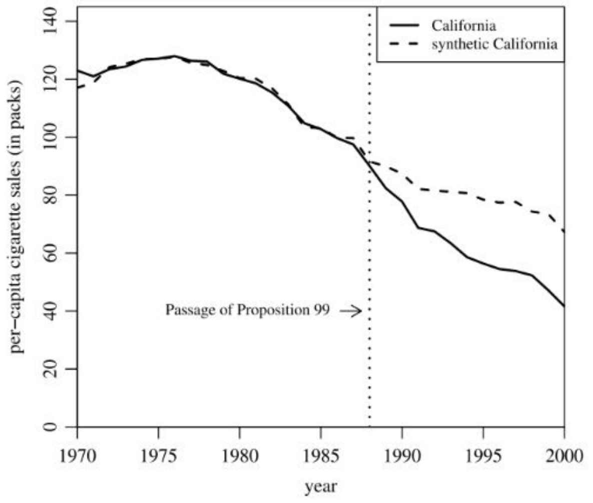


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

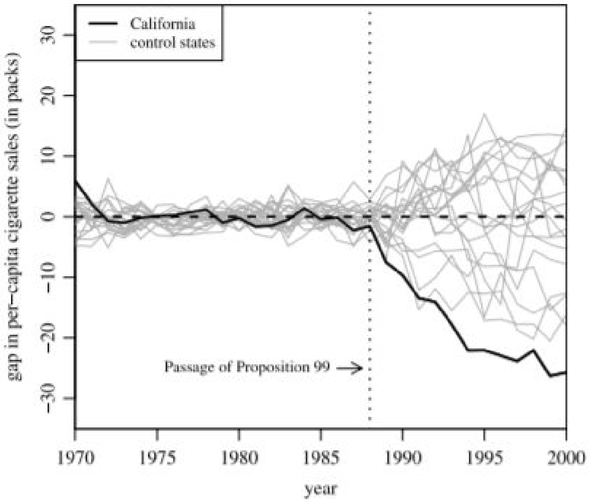


Figure 7. Per-capita cigarette sales gaps in California and placebo gaps in 19 control states (discards states with pre-Proposition 99 MSPE two times higher than California's).

Appendix C - Original Figures in the Basque Country Paper

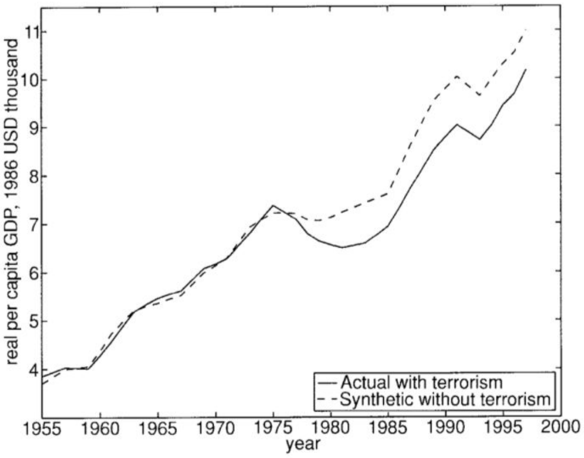


Figure 1. Per capita GDP for the Basque Country

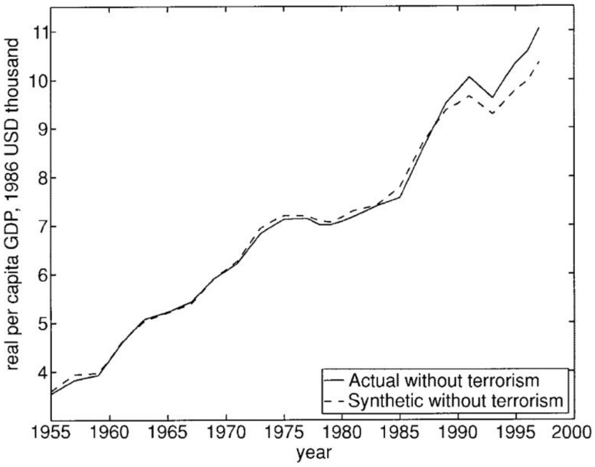
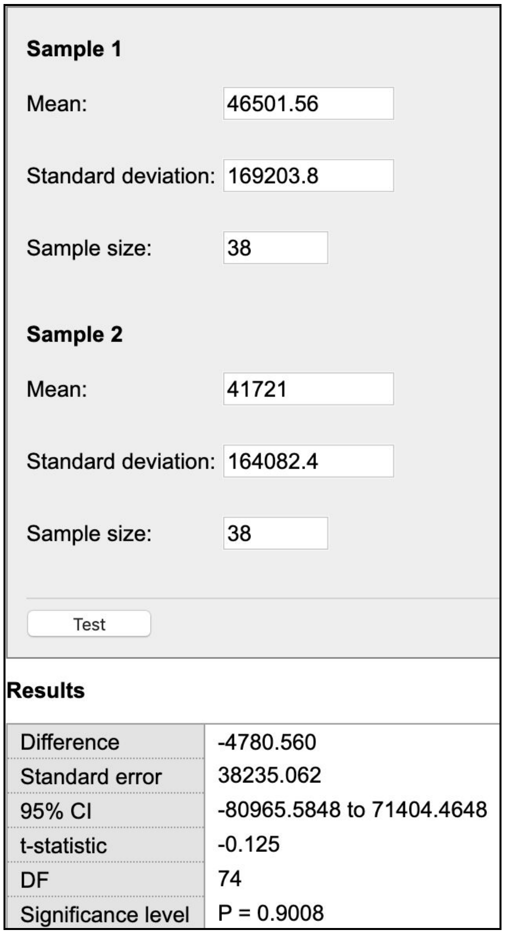


Figure 4. A "Placebo Study," per capita GDP for Catalonia

Appendix D - Significance Tests



Sample 1

Mean: 46501.56

Standard deviation: 169203.8

Sample size: 38

Sample 2

Mean: 41721

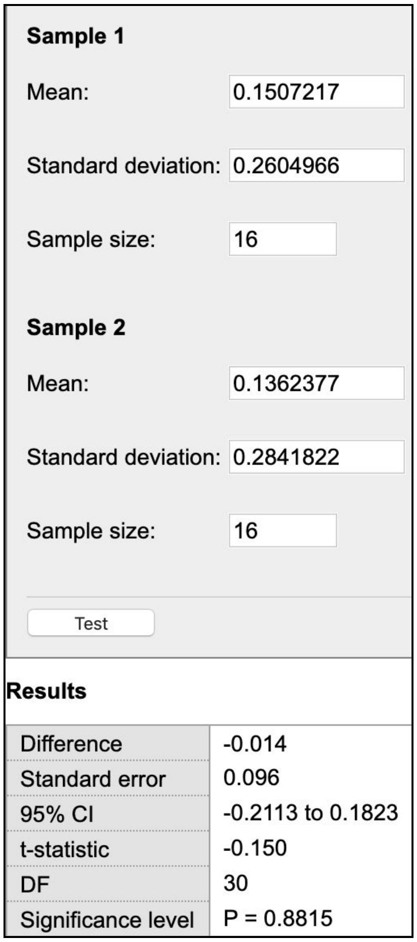
Standard deviation: 164082.4

Sample size: 38

Results

Difference	-4780.560
Standard error	38235.062
95% CI	-80965.5848 to 71404.4648
t-statistic	-0.125
DF	74
Significance level	P = 0.9008

Mean Squared Summed Error California



Sample 1

Mean: 0.1507217

Standard deviation: 0.2604966

Sample size: 16

Sample 2

Mean: 0.1362377

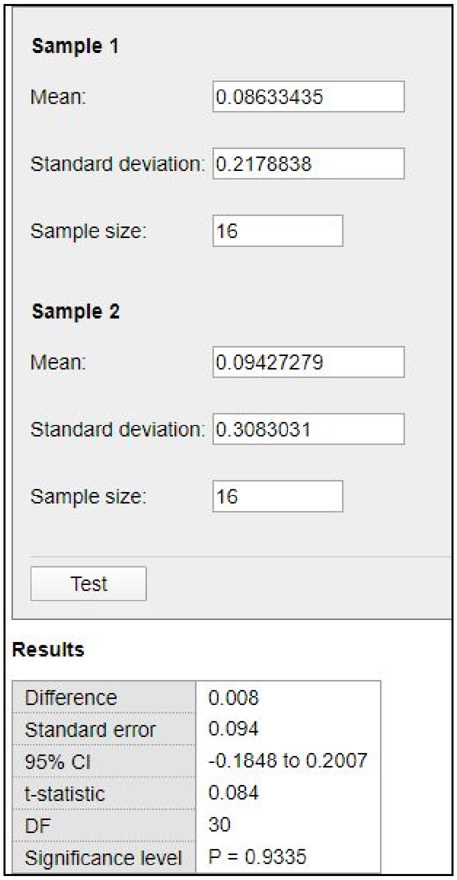
Standard deviation: 0.2841822

Sample size: 16

Results

Difference	-0.014
Standard error	0.096
95% CI	-0.2113 to 0.1823
t-statistic	-0.150
DF	30
Significance level	P = 0.8815

Mean Error Basque



Sample 1

Mean: 0.08633435

Standard deviation: 0.2178838

Sample size: 16

Sample 2

Mean: 0.09427279

Standard deviation: 0.3083031

Sample size: 16

Results

Difference	0.008
Standard error	0.094
95% CI	-0.1848 to 0.2007
t-statistic	0.084
DF	30
Significance level	P = 0.9335

Mean Squared Error Basque