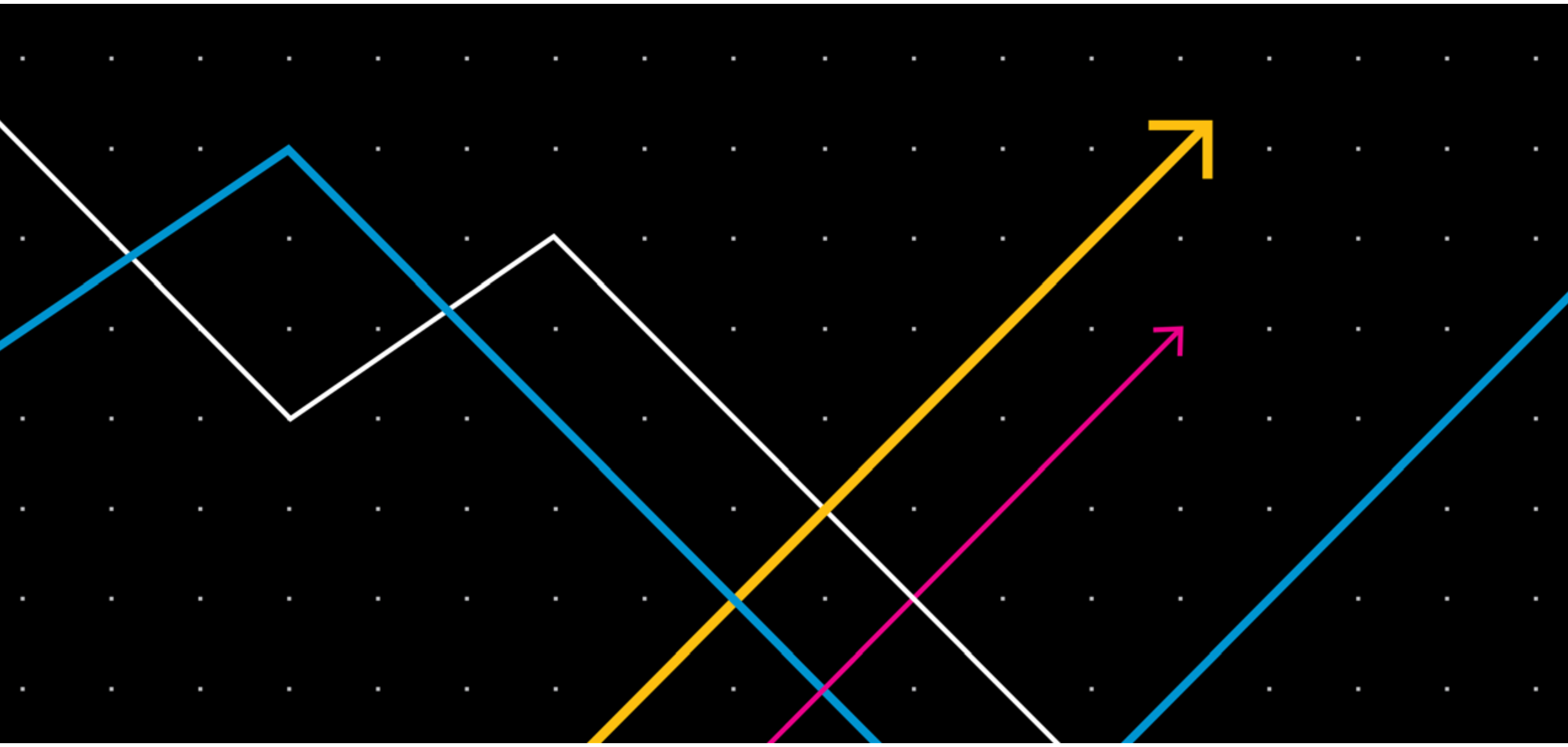
STATE AND LOCAL FINANCE INITIATIVE



RESEARCH REPORT

The Synthetic Control Method as a Tool to Understand State Policy

Robert McClelland *Sarah Gault*

March 2017

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

Cover image by Tim Meko.

Contents

Acknowledgments

This report was funded by the Laura and John Arnold Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at www.urban.org/support.

We gratefully acknowledge comments and suggestions from Norton Francis, Kim Rueben, and Rob Santos, and editing from the Urban Institute communications team.

The Synthetic Control Method as a Tool to Understand State Policy

Identifying good governance depends on accurately evaluating policies for their efficiency and effectiveness, and many tools are available for that purpose. Case studies of regional economies are an often-used example in which policies are evaluated through detailed examinations of economic conditions before and after the policies are implemented.

Because those analyses are qualitative rather than quantitative, rigorously identifying comparison groups whose outcomes can be contrasted with the outcome of the region undergoing the policy change is difficult. But without those control groups, separating the policy's effect from the effects of nonpolicy variables is nearly impossible. Nearby regions are sometimes used as controls for lack of alternatives, but geographic proximity is a poor metric for similarity if regions have substantial differences in political or cultural environments. Further, policies may spill across borders, confounding any comparison. The qualitative approach also impedes the analyst's attempt to generalize the analysis beyond the case at hand because few quantitative results can be applied to similar situations. Measures of similarity can be difficult to define, and no measures of statistical precision or accuracy exist.

An increasingly popular method for policy evaluation, the synthetic control method (SCM), addresses those problems. It provides quantitative support for case studies by creating a synthetic control region that simulates what the outcome path of a region would be if it did not undergo a particular policy intervention. The SCM creates this hypothetical counterfactual region by taking the weighted average of preintervention outcomes from selected donor regions. The donor regions that combine to form the synthetic control are selected from a pool of potential candidates. Predictor variables that affect the outcome and the outcome variable itself before the policy is enacted determine the selection of donor regions and weights. The resulting synthetic closely matches the affected region's outcome before policy enactment and is a control for the affected region following enactment. After policy enactment, the difference in outcomes between the affected region and its synthetic control counterpart reveals the policy's effectiveness.

The SCM is transparent. Analysts can evaluate how well the synthetic control's outcome matches the affected region's outcome before the policy change. In addition, donor regions and the weights assigned to them are known, and analysts can evaluate those regions' similarity to the policy region. It also does not require the same strict assumptions for accurate estimation as difference-in-differences or panel data methods. The researchers who popularized the method, Abadie, Diamond, and Hainmueller (2010)—hereafter, *ADH*—have made the source code publicly available, developing the SCM as a popular open-source modeling tool.

But ADH describe several assumptions necessary for the method to accurately estimate the policy change's effect. No region in the pool of potential donor regions can have a similar policy change. The policy in the affected region cannot affect the outcome in the pool of donor regions. To avoid possible interpolation bias, the variables used to form the weights must have values for the donor pool regions that are similar to those for the affected region. The values of those variables for the policy region cannot be outside any linear combination of the values for the donor pool. Finally, those variables and the outcome must have an approximate linear relationship.

In this report, we review the synthetic control method. We describe the method, using the ADH analysis as an example, and provide simple step-by-step guidance for using the method. We describe pitfalls and concerns noted in the literature, and we use the data from ADH to demonstrate these and other potential problems. In appendix A, we review the method's application in an economic development context in two recent articles. In using the SCM, these studies assess the "but for" test of causation ("but for" the intervention, the outcome change would not have happened) in economic development evaluation, particularly for public investment.

The Synthetic Control Method

The randomized controlled trial represents the gold standard for many studies of how a treatment can affect a particular outcome. In a common variation, people or entities (e.g., states or firms) are randomly drawn from a pool of similar candidates and placed in one of two groups. One group receives a treatment, and the second group does not. The treatment's effect is the difference between the outcome path for the treated group and the path for the control group. Although randomly treating whole regions is usually impossible, many natural experiments feature treatment variations across regions as if they were part of a randomized controlled trial. But that type of experiment is difficult to find when analyzing government policies because the jurisdictions enact the treatment, which is sometimes unique to that governmental entity. Aside from the selection problem, statistical methods such as difference-in-differences or fixed-effects estimation can be used to analyze natural experiments. Those methods, however, require that variables have fixed relationships over time.

The SCM addresses those problems by synthesizing a control from a weighted sum of donor regions chosen from a pool of potential candidates. That weighted sum is created by matching outcomes and explanatory variables in the pretreatment period of the donor regions to the same variables in the pretreatment period of the treated region. Valid implementation of the SCM requires that the control's outcome closely matches the treated entity's outcome during pretreatment. If the two match closely, comparing the outcome paths after the treatment provides insight about the treatment's effect. If the outcome paths of the synthetic control and the treated entity are similar in the treatment period, the treatment does not appear to have affected the outcome. If the paths diverge, however, the treatment presumably caused the difference.

To create a synthetic control, data must be available for several periods before the treatment for the policy region and the pool of potential donor regions. (For ease of exposition, we will refer to the regions as *states*.) Effective use of the method requires that three assumptions hold. First, only the treated state is affected by the policy change for all years in the pretreatment period used to create the synthetic control and afterward. Second, the policy change has no effect before it is enacted. And third, the treated state's counterfactual outcome can be approximated by a fixed combination of donor states.

Developed by Abadie and Gardeazabal (2003) to examine terrorism's effect on economic growth in Basque Country, the method became popular with other researchers following ADH. Its widespread use comes not only from the method's utility, but also from the source code's user availability. The *Synth* package, the software ADH developed to conduct the synthetic control analysis, is available on Hainmueller's webpage.[1] The package can be installed for MATLAB, Stata, and R. A sample dataset to replicate ADH's analysis is also available.[2] Galiani and Quistorff (2016) developed a complimentary package for Stata, *synth_runner*, to aid in running multiple estimations, assess the synthetic's fit, and graph the results.

ADH's Analysis of California's Tobacco Control Program

The SCM received much attention following ADH and the code's public release. In that article, the authors examine how California's tobacco control program under Proposition 99, implemented in 1988, affected smoking by creating a synthetic control version of California. They estimate that by 2000, per capita sales of cigarette packs had fallen 26 packs because of the program. Here, we describe the steps in analyzing a policy change with the SCM using the Proposition 99 example when possible. We also use ADH's original data to demonstrate the results' sensitivity to various modeling choices.

California voters approved Proposition 99 in 1988, an initiative that raised the cigarette excise tax by 25 cents a pack and implemented a large-scale antitobacco media campaign. The tax raised $100 million annually, and the revenue was initially directed toward antismoking efforts, including antismoking education budgets. Those efforts were substantially larger than efforts in other states, but the California assembly passed Assembly Bill 99 in 1991, which diverted a large share of the tax revenue for other purposes (Glantz and Balbach 2000). Beyond the tax increase, ADH report that Proposition 99 led to numerous local ordinances prohibiting smoking in indoor spaces such as restaurants and workplaces, and by 1993 almost two-thirds of employees in California worked in smoke-free environments. They also note that tobacco lobbyists in California responded by spending 10 times more in 1991 and 1992 than in 1985 and 1986.

After Proposition 99's apparent success, other states initiated similar policies. ADH state that Massachusetts raised taxes in 1993 to fund smoking-cessation efforts, and Arizona and Oregon raised taxes in 1994 and 1996, respectively. ADH also note that by 2009, 30 states and the District of Columbia required all workplaces, bars, and restaurants be smoke-free.

To determine Proposition 99's effectiveness in California (hereafter, *the treated state*), ADH focused on one outcome variable: per capita cigarette packs sold annually. ADH gathered data from 1970 through 2000. Because Proposition 99 was passed in 1988, ADH had 18 years of data before the policy was enacted and 12 years of data afterward. ADH ended their analysis period in 2000, citing an increasing number of state tobacco control initiatives in subsequent years that would have limited their pool of potential donor states. However, at the end of 1998 California voters approved Proposition 10, another tobacco control measure that implemented a cigarette excise tax hike of 50 cents a pack tax in 1999. To isolate the effects of Proposition 99 in California, ADH should have ended their sample period before Proposition 10 went into effect.

ADH assumed that the treatment had no effect on the outcome variable before the 1988 treatment (e.g., people did not stockpile cigarette packs in anticipation of the law's passage). That assumption, while convenient, is not critical for the SCM as long as it is clear when the effect started, because analysts can redefine the treatment to start at that point. ADH also assumed that the outcomes were independent across states, implying that the treatment in California did not affect the outcome in donor pool states. For example, they assumed that California residents did not respond to the tax increase by purchasing cigarette packs in nearby states. If that assumption fails to hold, the treatment can affect the donors to the synthetic control state, invalidating estimates of its effect drawn from the difference in outcomes between the treated state and the supposedly pristine control state.

Procedures

Box 1 provides a step-by-step overview of how to use the synthetic control method. Although each application will be different, these guidelines give practitioners a place to start. The analyst's first step is to identify predictors of the outcome variable. Ideally, those predictors have a stable relationship with the outcome variable. The predictors' ability to explain variation over the pretreatment years, however, is less important because only their time averages over pretreatment years are used when creating the synthetic state. For example, ADH used the log of per capita GDP averaged over 1980–88, the share of the population between ages 15 and 24 averaged over 1980–88, per capita beer consumption averaged over 1984–88, and the average retail price of cigarettes averaged over 1980–88. Hahn and Shi (2016) state that studies should have a relatively large number of predictors compared to the number of donor states. That may seem counterintuitive to analysts used to worrying about degrees of freedom, but Hahn and Shi (2016) point out that in the SCM it improves the choice of weights assigned to each state.

One potentially important predictor of the outcome variable is the lagged outcome variable. It avoids the problem of omitting important predictors' effects because it includes the effects of any predictor variables whether or not they are gathered by the analyst. Including the lagged outcome variable for some pretreatment years is common, and Athey and Imbens (2006) even state that including the other covariates rarely matter. An analyst might be tempted to use the values of the outcome variable for all past years. But Kaul and coauthors (2016) show that doing so eliminates all other predictors' effects, which may account for why other covariates have no apparent effect in some studies. That is, the same synthetic control state is created regardless of the other predictors' values. Kaul and coauthors (2016) also point out that if the other predictors help predict the outcome, omitting them can bias the synthetic regions' outcome in the posttreatment period. They suggest using either an average of the outcome across all pretreatment years or the last year of the pretreatment period. Ferman, Pinto, and Possebom (2016) recommend that analysts try several different sets of lags and report the results from all of them. ADH use three years of per capita annual sales of cigarette packs: 1975, 1980, and 1988. Based on an analysis of the ADH data discussed below, we recommend that analysts choose a small number of lags that follow the outcome trend in the pretreatment period.

The next step is to identify the potential donor states that will synthesize the control state. Because the control state is a contrast to the treated state after treatment, similar policies should not be enacted in any donor pool state in any year during the study. ADH dropped four states that introduced tobacco control programs during the treatment period (Arizona, Florida, Massachusetts, and Oregon) and seven states that raised cigarette taxes by at least 50 cents a pack during that period (Alaska, Hawaii, Maryland, Michigan, New Jersey, New York, and Washington). ADH noted that using those excluded states could have reduced the difference between the synthetic control state and treated state outcomes. If states with a similar treatment had been included, the resulting difference would be a lower bound on the policy's effect. Had they existed, states with similar treatments between 1970 and 1988 also should have been eliminated. Finally, ADH eliminated the District of Columbia, which left 38 possible donor states. The relationship between the predictors and the outcome variable in the donor pool states must be similar to that relationship in the treated state. If the relationship in the donor states is substantially different, the resulting synthetic control state will not effectively represent a counterfactual to the treated state. In the next section, we explore how ADH's results are affected when fewer predictors are used.

The SCM must approximate the treated state under the counterfactual assumption that the treatment did not occur. The counterfactual state must be accurately represented by a fixed combination of donor states. That would be violated if, for example, the share of Californians between ages 15 and 24 who smoked was much lower than in potential donor states or if the price of cigarettes in California was substantially higher than in potential donor states. There must be states with predictor values near those of the treated state before treatment. ADH illustrate this with a problematic example: if one explanatory variable is the share of the population that is white, and California is 80 percent white, it could be represented by a synthetic state formed from an equal weighting of a 65 percent white/35 percent nonwhite state with a 95 percent white/5 percent nonwhite state. But if a strong nonlinear relationship exists between racial composition and smoking, the synthetic state could poorly approximate the treated state, resulting in interpolation bias. Abadie, Diamond, and Hainmueller (2015) also declare that the outcomes of the states in the donor pool should be driven by the same process as that found in the treated state before treatment. Enlarging the pool by including states with idiosyncratic variation in predictors runs the risk of overfitting, in which case the resulting synthetic control might poorly mimic the outcome of the treated state in the absence of the treatment.

The SCM creates state-level weights to form a synthetic control state. The weights, which must be nonnegative and sum to one, are contingent on other weights that indicate the power of the chosen predictor variables. ADH state that the resulting analysis is valid for any predictor weights, and they suggest several possibilities for selecting the weights. One option is to subjectively choose the weights, bypassing econometric approaches. In their model, ADH choose weights from all possible matrices to minimize the outcome's mean squared prediction error (MSPE) in the pretreatment years. A third option described in Abadie, Diamond, and Hainmueller (2015) is designed to reduce overfitting through a form of cross-validation in which the pretreatment period is divided into a training period and a validation period. However, Klößner and coauthors (2016) point out that the results are sensitive to the number of predictors and selected donor states. The default method in the *Synth* package for selecting the predictor weights is another regression-based method described in Kaul et al. (2016) (appendix B). ADH note that the MSPE of the synthetic is affected by the chosen weights, so optimal weights will minimize that MSPE and provide the best fit.

BOX 1

The Synthetic Control Method

Analysts can use the synthetic control method to study a policy treatment's effect in a state on a particular outcome by comparing the treatment outcome with that of the synthetic control. After determining a period of interest (often based on the availability of data on the outcome variable) the steps to generate the synthetic control are as follows:

Step 1: Identify predictors of the outcome variable.

- Choose predictor variables that should affect outcomes in states both before and after treatment.
- Determine the pretreatment year range over which the predictors will be averaged. A longer pretreatment year range is better than a short one, and variables do not need to be rejected if some years in the pretreatment period are unavailable.
- Include several (but not all) lagged values of the outcome variable with the other predictors. Choose values that highlight the trend of the outcome before treatment.

Step 2: Identify possible donor states to synthesize the control state.

- Exclude any states that enacted policy treatments of similar or larger size during the selected period. Relatively small treatments do not necessarily disqualify a state.
- States in the donor pool should have values of predictors that are close to the values of the treated state before treatment. The values of each predictor in the treated state must be neither the largest nor smallest and ideally lie towards the middle.

Step 3: Choose a method for selecting predictor weights.

- The optimal weights will minimize the synthetic's mean squared prediction error.
- The cross-validation method shows promise, but currently the standard method is the safer choice.

Step 4: Assess the pretreatment period goodness of fit of the synthetic control state (generated using the *Synth* package).

- Evaluate how closely the outcome path of synthetic control during the pretreatment period follows that of the treated state based on appearance and the root mean squared prediction error.
- If the fit appears poor, use a model with all possible outcome lags as a test. If the synthetic control state under this model poorly matches the treated state, the synthetic control method should not be used because no model will yield a good fit. But recognize that using all possible lags in the final model can bias the outcome path of the synthetic.
- Review state weights to judge similarities between the donor states and the treated state. It can be more important for outcomes of donor states to have a trend similar to that of the treated state than for the states to have a similar average.
- Review predictor weights to determine the selected predictor variables' strength in explaining the outcome.

Step 5: Conduct placebo test on states in the donor pool to evaluate the significance of the results for the treated state.

- If the posttreatment difference between the treated state and its synthetic is larger than the difference for most of the placebo states, there is evidence that the treatment had an effect.
- Evidence of significance should be treated as suggestive of an effect rather than as a rejection of a null hypothesis.

Step 6: Conduct sensitivity analyses to further test the credibility of the results.

Output

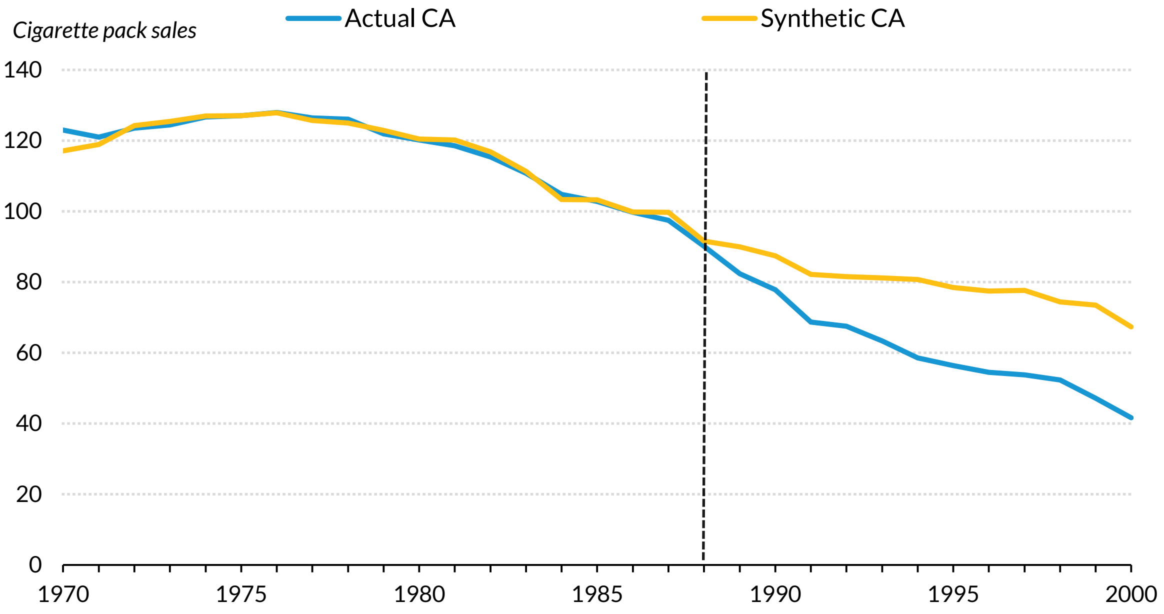
The SCM's primary output is a pretreatment and posttreatment path for the synthetic control state's outcome variable that can be compared with the treated state's outcome variable path. Ideally, the two paths follow each other closely before the treatment, so that divergence after that point can represent the treatment's effect. Analysts can assess the goodness of fit by calculating the root mean squared prediction error (RMSPE) between the actual and synthetic region during the pretreatment period or by using the "eyeball test," though appearance alone is not always sufficient. A poor fit could be caused by several factors, such as using weak predictors, using outcome variables from problematic pretreatment years as predictors, or using predictors for which the treated state has extreme values relative to the donor states.

Figure 1 shows results replicated from ADH, in which the synthetic California was created by minimizing the MSPE between the actual and synthetic California over the pretreatment years (1970–88). Annual cigarette sales of packs per capita in the synthetic California closely follow the sales in the actual California until 1988, meaning that the synthetic California appears to be a control for the actual California. After 1988, the curves diverge, and by 2000, the gap exceeds 26 packs annually per capita.

FIGURE 1

Synthetic California Per Capita Cigarette Sales (ADH)

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from ADH to replicate ADH's analysis.
Note: ADH = Abadie, Diamond, and Hainmueller (2010).

The predictors of cigarette sales—which include lagged values of cigarette sales and the time averages of nonlag predictors—closely match for the synthetic and actual California in the years leading up to Proposition 99's passage (table 1), contributing to the close pretreatment fit in ADH's model. But the synthetic region's pretreatment outcome might not tightly fit the actual region's outcome. In either case, the list of donor regions and their associated weights should be published so that the analyst and subsequent readers can judge the donor regions' similarity to the treatment region. Table 2 lists the selected donor states and associated weights replicated from ADH. Of the five states, Utah is the most heavily weighted (33.5 percent). Nevada and Montana, respectively, receive weights of 23.5 and 20.1 percent. Colorado has a weight of 16.1 percent, and Connecticut has the smallest weight at 6.8 percent. It might seem strange that Utah and Nevada receive the largest weights while Connecticut, which might be culturally closer to California, receives the smallest. Plots of cigarette sales for those three states in figure 2 make the relative weights even more difficult to understand. Annual cigarette sales per capita in Connecticut are far closer to California's than they are to Utah's or Nevada's.

TABLE 1

Actual and Synthetic California Predictor Means (ADH)

Variable	Years	Actual CA	Synthetic CA
Beer consumption per capita	1984–88	24.28	24.21
Log state per capita GDP	1980–88	10.08	9.86
Retail price of cigarettes	1980–88	89.42	89.41
Share of state population ages 15–24	1980–88	0.17	0.17
Cigarette sales per capita, 1988	1988	90.1	91.64
Cigarette sales per capita, 1980	1980	120.2	120.45
Cigarette sales per capita, 1975	1975	127.1	127.06

Source: Authors' calculations using the *Synth* package and data from ADH to replicate ADH's analysis.
Notes: ADH = Abadie, Diamond, and Hainmueller (2010). Units are gallons for beer consumption per capita, cents for retail price of cigarettes, and packs for cigarette sales per capita.

TABLE 2

Synthetic California Donor State Weights (ADH)

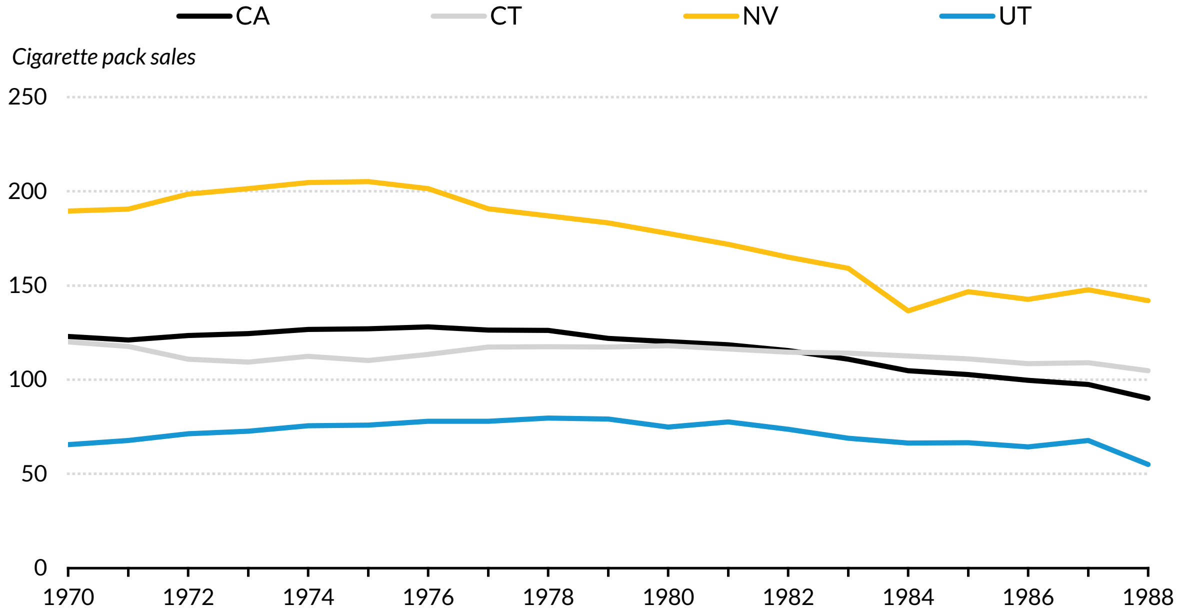
State	Weight
Colorado	0.161
Connecticut	0.068
Montana	0.201
Nevada	0.235
Utah	0.335
Sum	**1.000**

Source: Authors' calculations using the *Synth* package and data from ADH to replicate ADH's analysis.
Note: ADH = Abadie, Diamond, and Hainmueller (2010).

FIGURE 2

Per Capita Cigarette Sales in California and Selected Donor States

Before Proposition 99 passage in 1988



Source: Authors' analysis of data from Abadie, Diamond, and Hainmueller (2010).

Examining the predictor weights sheds light on this question. Table 3 lists the weights replicated from ADH. The cigarette-sales-in-1975 lagged outcome variable receives more than 80 percent of the weight. When fitting the path of cigarette sales in the synthetic California to the path in the actual California, the SCM selects states based mostly on their sales in 1975. This suggests that cigarette sales in the synthetic California will most closely match sales in the actual California in 1975 and that matches in other years will depend on how closely sales in the donor states follow the path taken by California's sales. To demonstrate how the reliance on 1975 affects donor state weights, we plot cigarette sales in California, Connecticut, Nevada, and Utah after indexing their sales so each state's sales equals 1 in 1975 (figure 3). That forces the value of sales in each state to converge in 1975 (which is what would happen if sales in 1975 were the only variable used to match donor states to California). The paths of Nevada and Utah sales bracket California's path, while the path taken by Connecticut sales is much flatter and drifts away from California. From figure 3, we can understand how a strong reliance on cigarette sales in 1975 leads to large state weights on Nevada and Utah in the ADH model.

TABLE 3

Synthetic California Predictor Weights (ADH)

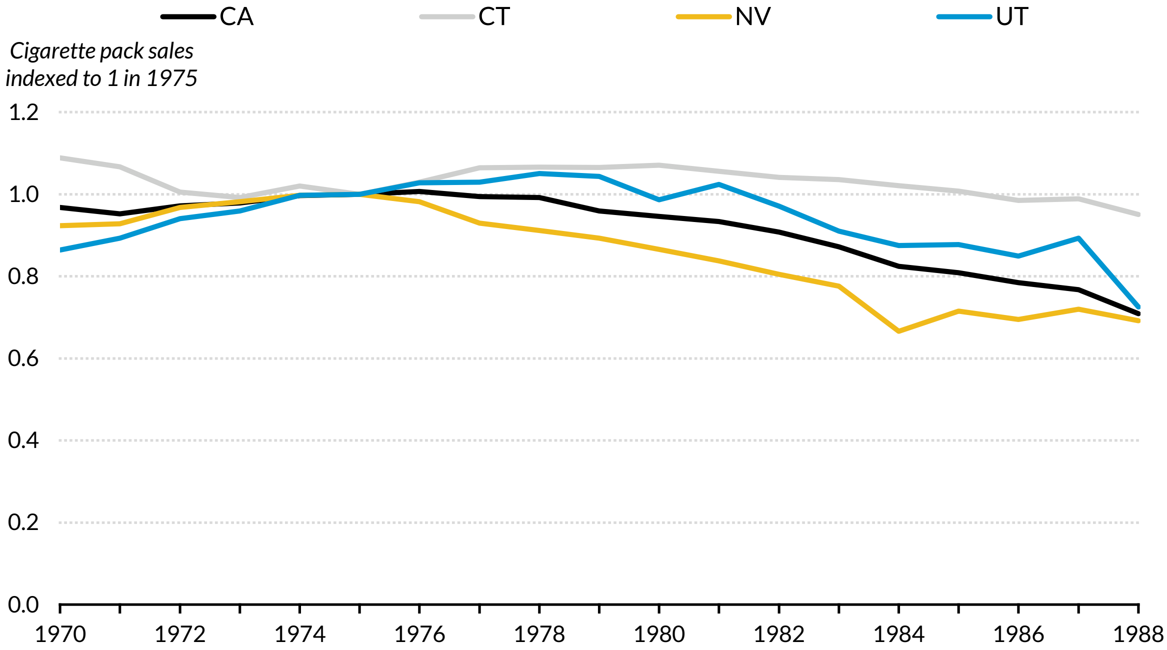
Variable	Weight
Beer consumption per capita	0.013
Log state per capita GDP	0.0002
Retail price of cigarettes	0.124
Share of state population ages 15–24	0.016
Cigarette sales per capita, 1988	0.024
Cigarette sales per capita, 1980	0.011
Cigarette sales per capita, 1975	0.811
Sum	**1.000**

Source: Authors' calculations using the *Synth* package and data from ADH to replicate ADH's analysis.
Note: ADH = Abadie, Diamond, and Hainmueller (2010).

FIGURE 3

Per Capita Cigarette Sales in California and Selected Donor States Indexed to 1 in 1975

Before Proposition 99 passage in 1988



Source: Authors' analysis of data from Abadie, Diamond, and Hainmueller (2010).

Confidence intervals and tests of significance are not calculated in the SCM, leaving analysts without the standard methods for calculating the statistical significance of the posttreatment outcome difference between the treated and synthetic states. That deficiency can be addressed using the placebo tests described in ADH. In those tests, the SCM is separately run on each state in the donor pool as though it is a treated state, using the remaining members of the pool as before. The resulting placebo state is compared with its synthetic match, and the test is repeated on the next state in the donor pool. Because none of the donor pool states receive treatment, variation between the placebo state and its synthetic match occurs randomly. By comparing the difference between the treated state and its synthetic control to the differences among placebo states and their controls, we can evaluate the likelihood that the treatment's apparent effect on the treated state is because of chance. ADH and Abadie, Diamond, and Hainmueller (2015) also use placebo runs to compare the difference in the synthetic control's RMSPE before and after treatment for the treated state and each placebo state. If the treatment is effective, after the intervention the treated state's actual path will diverge away from the synthetic control. The RMSPE of the treated state's synthetic control after the treatment then will be large relative to its value before treatment. Placebo states, on the other hand, should not see a substantial increase in their RMSPE following the treatment.

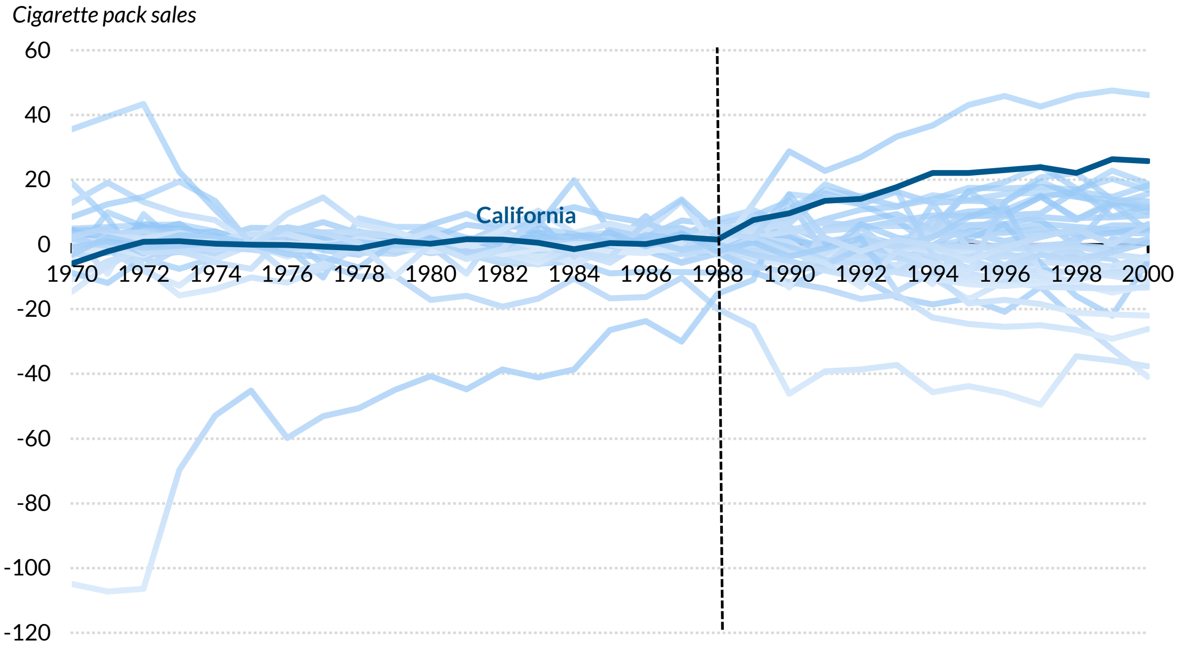
However, analyses in recent work by Ferman and Pinto (2016) and Hahn and Shi (2016) have questioned whether the placebo tests are useful for hypothesis testing. For example, Ferman and Pinto (2016) note that while analysts only use the SCM if the synthetic control closely matches the treated state during the pretreatment period, the same condition does not always hold for each of the placebo states. Yet the comparison for treated and placebo states is not valid unless it does hold. Nevertheless, a graphical comparison of the actual-to-synthetic difference for the treated and placebo states illuminates how the path of the treated state differs from those of the placebo states. It therefore can still provide useful information about the potential effectiveness of the treatment.

Here, we used the ADH data to construct the difference between the synthetic control state and the treated state for California and 37 placebos. (Because Utah has the lowest per capita cigarette sales in each year, it is not possible to create its synthetic control.) The results demonstrate that before 1988, the gap between California and its synthetic control is smaller than a similar gap for most placebo states (figure 4). The line far below the others in the pretreatment period represents New Hampshire. ADH state that New Hampshire's poor match is unsurprising because that state has the highest per capita cigarette sales in each year before 1988. A weighted sum of the sales from the remaining states cannot equal New Hampshire's.[3] After 1988, the gap between the per capita sales of the synthetic and actual California is larger in magnitude than the gap for most placebo states. The small gap before 1988 and the large and growing gap after 1988 strongly suggest that California's gap is because of Proposition 99. ADH conduct an alternative version of the placebo test where they exclude five placebo states with pretreatment MSPEs over 20 times that of the synthetic California (further versions exclude states with MSPEs five times or twice the size of synthetic California's). Although the SCM is problematic for these cases with a poor pretreatment fit, we include them in figure 4 to show the full distribution of placebo state synthetics.

FIGURE 4

Difference in Synthetic and Actual Per Capita Cigarette Sales for California and 37 Placebo States

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).
Note: The *Synth* package cannot generate a synthetic placebo for Utah, one of the 38 donor pool states.

Sensitivity Tests

We conducted sensitivity tests to assess the potential pitfalls noted by ADH and others, using the data from ADH's study of the effect of California's Proposition 99 on smoking and their *Synth* package.[4] We evaluate the effects of the choice of outcome lags used as predictors, the choice of other predictors, the length of the pretreatment year range used to fit the synthetic region, and the method for choosing predictor weights. We use three criteria: sensitivity of the fit between the synthetic state and the treated state outcomes in the pretreatment period, sensitivity of the synthetic state outcome in the treatment period, and sensitivity of donor state selection. Table 4 shows RMSPEs—one measure of the pretreatment fit—for the different model choices we analyze. We conduct two additional sensitivity tests described in Abadie, Diamond, and Hainmueller (2015)—the in-time placebo test and the leave-one-out test—to assess the validity of the results.

TABLE 4

Synthetic California Root Mean Squared Prediction Error

	RMSPE
Outcome lags	
1988	4.197
1975	7.918
1980 & 1988	4.499
1975 & 1988	1.754
1975, 1980, & 1988 (ADH)	1.757
1970–88 all	1.656
1970–88 average	4.709
Predictor variables	
Four predictors + lags (ADH)	1.757
Three predictors + lags	2.950
Two predictors + lags	3.420
One predictor + lags	2.437
Lags only	4.329
Four predictors, no lags	16.328
Three predictors, no lags	10.740
Two predictors, no lags	9.131
Predictor year range	
1970–88	1.730
1975–88	1.731
1980–88 (ADH)	1.757
1984–88	1.924
Method for selecting predictor weights	
Standard	1.740
Cross-validation	2.373

Source: Authors' calculations using the *Synth* package and data from ADH.
Notes: ADH = Abadie, Diamond, and Hainmueller (2010). RMSPE = root mean squared prediction error. RMSPEs are measured over the full 1970–88 pretreatment period with the exception of the cross-validation method RMSPE, which is measured over the 1980–88 validation period.

Outcome Lags

We first examine how different outcome lags affect the analysis. ADH include three lags of their outcome variable, per capita cigarette sales, for 1975, 1980, and 1988 with their other predictor variables. We compare ADH's original lag selection with combinations ranging from one lag (1975 or 1988) to all lags. We examine Kaul and coauthors' (2016) concern that including all outcome lags for all pretreatment years will render the other predictor variables (averaged over pretreatment years) useless. We also include the two alternatives they put forth: using an average of the outcome variable across all pretreatment years or using only the lag for the final year of the pretreatment period (1988).

The pretreatment outcomes vary noticeably with the choice of outcome lags. Unsurprisingly, the synthetic control California that includes all outcome lags as predictor variables closely matches the actual California in the pretreatment period (figure 5). Using the three lags chosen by ADH fits almost as well, although the RMSPE is about 6 percent higher (table 4). When we include all lags in the model, other predictors have no effect on the pretreatment outcome for the synthetic California. Including all outcome lags results in a close-fitting synthetic in the pretreatment period but eliminates the benefit from including any other predictors. Without any other predictor variables, no theory behind the model suggests what affects future cigarette sales (other than past sales). That result confirms Kaul and coauthors' (2016) analysis, but the synthetic California using the average-lag or final-year-lag outcomes—their preferred alternatives—fit more poorly than most other alternatives in visible fit and RMSPE.

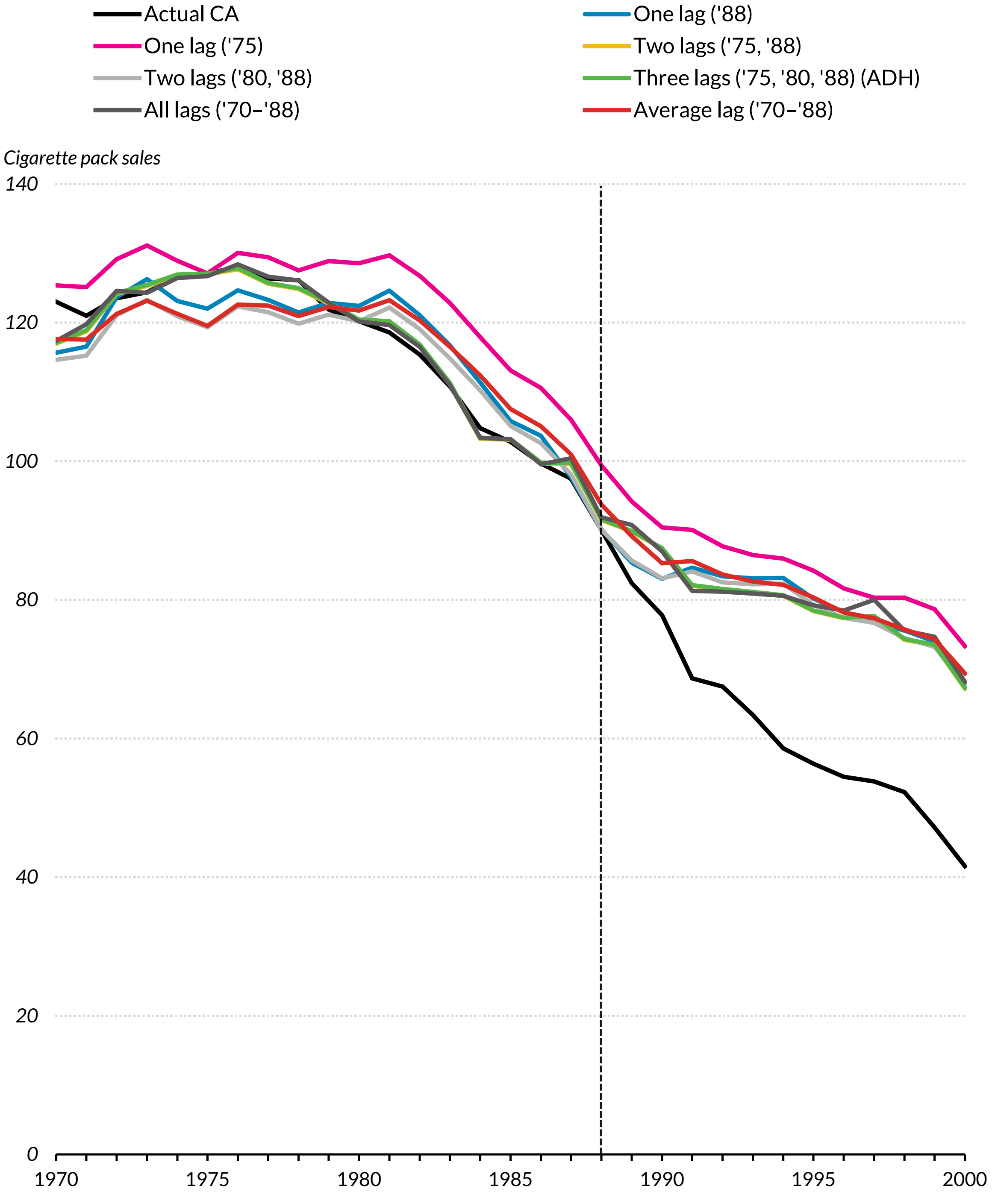
Adding the 1975 outcome to the final-year-lag model creates a closer-fitting synthetic California in the pretreatment period in RMSPE and visual fit. But the year is important: adding the 1980 lag instead of 1975 worsens the fit, as the RMSPE jumps from 1.754 to 4.499. Yet while the 1975 outcome seems to play an important role, using only that lag produces the worst fit of all our lag tests, visually and for the RMSPE. The outcome path for this synthetic with only one lag for 1975 is also noticeably higher than the paths of the other synthetics in the posttreatment years.

The choice of output lags plays a large role in the model's fit because it influences the states used. Table 5 lists the state weights for the lag tests in figure 5. In each instance where 1975 and at least one other year is used, Montana, Nevada, and Utah each receive large weights, and the synthetic California closely matches the actual California in the pretreatment period. In these cases, the RMSPE is below 1.8. Figures 2 and 3 demonstrate why the RSMPE is low when Nevada and Utah are used, as their cigarette sales paths bracket California's. In all other instances listed, the synthetic California poorly matches the actual California, as Montana, Nevada, and Utah receive small or zero weights. Instead, Colorado or Connecticut receive large weights, and the RMSPE exceeds 4.1 in these cases. Figures 2 and 3 also demonstrate why the RMSPE is so high when Connecticut receives a relatively large weight: Connecticut's cigarette sales levels are similar to California's over the pretreatment period (figure 2), but they do not follow the same path (figure 3). Notably, the lags Kaul and coauthors (2016) recommend in place of the problematic all-lag model do not result in a synthetic California that closely matches the actual California in the pretreatment period.

FIGURE 5

Synthetic California Per Capita Cigarette Sales with Various Outcome Lags

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from ADH.

Notes: ADH = Abadie, Diamond, and Hainmueller (2010).

TABLE 5

Synthetic California State Weights for Various Outcome Lags

	1988	1975, 1988	1975	1980, 1988	1975, 1980, 1988 (ADH)	All years 1970–88	Average 1970–88
Alabama	0	0	0	0	0	0	0
Arkansas	0	0	0	0	0	0	0
Colorado	0.733	0.16	0.651	0.664	0.161	0.015	0.609
Connecticut	0.064	0.068	0.300	0.159	0.068	0.109	0.294
Delaware	0	0	0	0	0	0	0
Georgia	0	0	0	0	0	0	0
Idaho	0	0	0	0.03	0	0	0
Illinois	0	0	0	0	0	0	0
Indiana	0	0	0	0	0	0	0
Iowa	0	0	0	0	0	0	0
Kansas	0	0	0	0	0	0	0
Kentucky	0	0	0	0	0	0	0
Louisiana	0	0	0	0	0	0	0
Maine	0	0	0	0	0	0	0
Minnesota	0	0	0	0	0	0	0
Mississippi	0	0	0	0	0	0	0
Missouri	0	0	0	0	0	0	0
Montana	0	0.201	0	0	0.201	0.232	0
Nebraska	0	0	0	0	0	0	0
Nevada	0	0.235	0.035	0	0.235	0.205	0
New Hampshire	0	0	0	0	0	0.045	0
New Mexico	0.128	0	0	0	0	0	0
North Carolina	0	0	0	0	0	0	0
North Dakota	0	0	0	0	0	0	0
Ohio	0	0	0	0	0	0	0
Oklahoma	0	0	0	0	0	0	0
Pennsylvania	0	0	0	0	0	0	0
Rhode Island	0	0	0	0	0	0	0
South Carolina	0	0	0	0	0	0	0
South Dakota	0	0	0	0	0	0	0
Tennessee	0	0	0	0	0	0	0
Texas	0	0	0	0	0	0	0
Utah	0.076	0.335	0	0.147	0.335	0.394	0.097
Vermont	0	0	0	0	0	0	0
Virginia	0	0	0	0	0	0	0
West Virginia	0	0	0	0	0	0	0
Wisconsin	0	0	0.014	0	0	0	0
Wyoming	0	0	0	0	0	0	0
Sum	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Source: Authors' calculations using the *Synth* package and data from ADH.
Note: ADH = Abadie, Diamond, and Hainmueller (2010).

The choice of outcome lags plays an important role in selecting the regions used to form the synthetic region. Comparing the synthetic region's fit in several specifications and examining the corresponding change in selected donor regions and their weights allows analysts to improve the match between the synthetic and treated regions.

Nonlag Predictor Variables

We examine the importance of predictors other than outcome lags with the four predictor variables (averaged over pretreatment years) used by ADH. In order of their predictive power in ADH's model, these variables are as follows: cigarettes' average retail price, the share of the population ages 15 to 24, per capita beer consumption, and the log of per capita GDP. We remove the predictor with the lowest weight from the model and continue until we are left with only the three outcome lags used by ADH (table 6).

TABLE 6

Predictor Variables Included in the Sensitivity Test Subsets

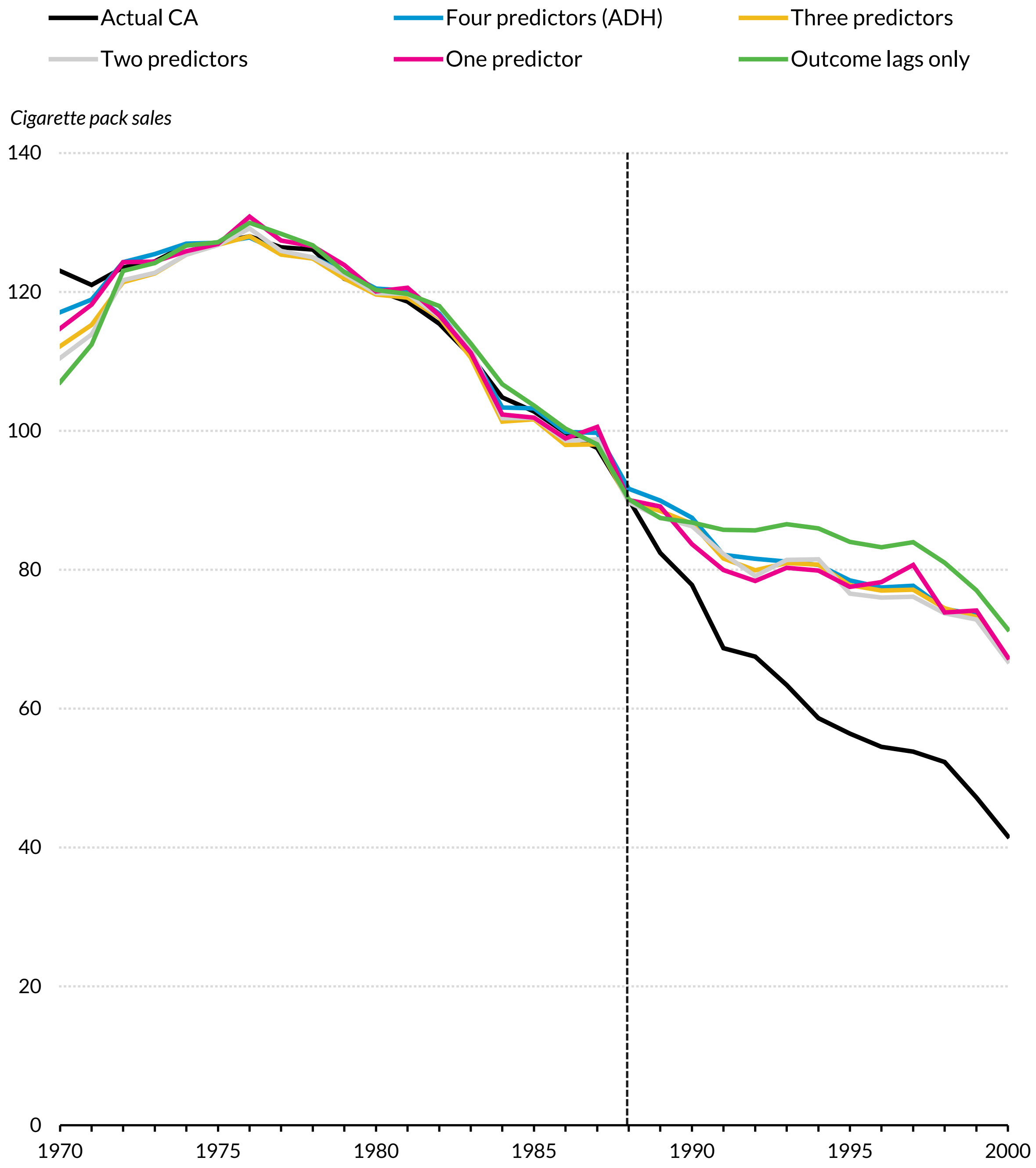
	Four predictors (ADH)	Three predictors	Two predictors	One predictor	Outcome lags only
Log state per capita GDP	X				
Beer consumption per capita	X	X			
Share of state population ages 15–24	X	X	X		
Retail price of cigarettes	X	X	X	X	
Cigarette sales per capita, 1988	X	X	X	X	X
Cigarette sales per capita, 1980	X	X	X	X	X
Cigarette sales per capita, 1975	X	X	X	X	X

As predictors are removed (figure 6), the synthetic California begins to stray from the actual California for the first five pretreatment years (1970–75). Beyond 1975, the pretreatment outcomes appear to be similar, which might suggest that using only the price of cigarettes as a predictor is sufficient. But the RMSPEs vary markedly from one model to the next. From table 4, using all four predictors and the three outcome lags has the lowest RMSPE: 1.757. Removing the variable with even the lowest predictive power increases the RMSPE to nearly three. Removing a second increases the RMSPE further, but removing a third decreases the RMSPE. Removing variables does not automatically worsen the fit, and adding variables does not necessarily improve the fit. Removing the final nonlag predictor so only the three outcome lags are used increases the RMSPE past four. The outcome path of this synthetic California with no predictor variables other than the 1975, 1980, and 1988 outcome lags is higher than that of the other synthetic Californias in the years following the treatment.

FIGURE 6

Synthetic California Per Capita Cigarette Sales with Various Predictor Variables (in Addition to Outcome Lags)

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from ADH.

Note: ADH = Abadie, Diamond, and Hainmueller (2010).

TABLE 7

Synthetic California State Weights for Various Predictor Variables (in Addition to Outcome Lags)

	Four				**Outcome**	**Four, no**	**Three, no**	**Two, no**
	(ADH)	**Three**	**Two**	**One**	**lags only**	**lags**	**lags**	**lags**
Alabama	0	0	0	0.001	0.003	0	0.023	0.024
Arkansas	0	0	0	0.001	0.003	0	0.02	0.021
Colorado	0.161	0	0.001	0.001	0.021	0	0.019	0.019
Connecticut	0.068	0	0.001	0.003	0.003	0.459	0.049	0.092
Delaware	0	0	0	0.001	0.002	0.147	0.029	0.026
Georgia	0	0	0	0.001	0.003	0	0.021	0.021
Idaho	0	0.069	0.185	0	0.215	0	0.016	0.017
Illinois	0	0	0.001	0.001	0.005	0	0.022	0.023
Indiana	0	0	0	0	0.002	0	0.018	0.018
Iowa	0	0	0.001	0.001	0.005	0	0.023	0.024
Kansas	0	0	0.001	0.001	0.005	0	0.02	0.022
Kentucky	0	0	0	0	0.002	0	0.015	0.013
Louisiana	0	0	0.001	0.001	0.005	0	0.035	0.028
Maine	0	0	0	0.001	0.003	0	0.022	0.023
Minnesota	0	0.015	0.002	0.004	0.005	0	0.037	0.032
Mississippi	0	0	0	0.001	0.003	0	0.029	0.026
Missouri	0	0	0	0.001	0.003	0	0.018	0.019
Montana	0.201	0.193	0	0.001	0.187	0	0.017	0.019
Nebraska	0	0	0.001	0.001	0.005	0	0.021	0.022
Nevada	0.235	0.249	0.218	0.177	0.002	0	0.009	0.023
New Hampshire	0	0	0.022	0.111	0.008	0	0.014	0.018
New Mexico	0	0.005	0.001	0.001	0.01	0	0.021	0.022
North Carolina	0	0	0	0	0.135	0	0.016	0.014
North Dakota	0	0.136	0.219	0.119	0.043	0	0.029	0.026
Ohio	0	0	0	0.001	0.003	0	0.019	0.019
Oklahoma	0	0	0.002	0.004	0.009	0	0.021	0.023
Pennsylvania	0	0	0	0.001	0.003	0	0.019	0.021
Rhode Island	0	0	0	0.001	0.002	0	0.026	0.025
South Carolina	0	0	0	0	0.003	0	0.019	0.017
South Dakota	0	0	0.001	0.001	0.005	0	0.022	0.023
Tennessee	0	0	0	0.001	0.002	0	0.019	0.02
Texas	0	0	0.002	0.003	0.006	0	0.037	0.029
Utah	0.335	0.331	0.337	0.554	0.258	0	0.046	0.119
Vermont	0	0	0	0.001	0.004	0	0.024	0.024
Virginia	0	0	0	0	0.003	0.394	0.016	0.015
West Virginia	0	0	0	0.001	0.003	0	0.021	0.023
Wisconsin	0	0	0.001	0.001	0.004	0	0.149	0.031
Wyoming	0	0	0	0.001	0.017	0	0.018	0.018
Sum	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Source: Authors' calculations using the *Synth* package and data from ADH.
Note: ADH = Abadie, Diamond, and Hainmueller (2010).

Changing the number of predictors changes the donor states selection more dramatically than changing the outcome lags (table 7). Dropping the least important predictor (the log of per capita GDP) drops Colorado and Connecticut from the selected states and adds four new donor states. Dropping additional predictors further increases the number of selected states, although Montana, Nevada, and

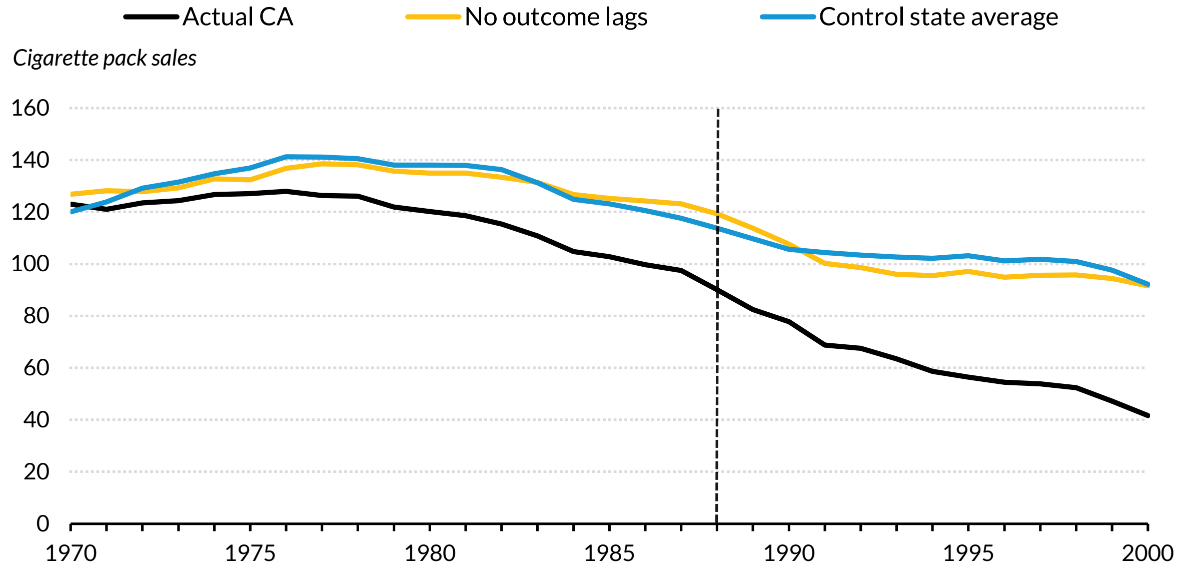
Utah continue to receive substantial weights. Using only one predictor leads to a positive weight for 32 states, although Utah and Montana are heavily weighted and New Hampshire and North Dakota receive smaller but notable weights. Using only the three outcome lags from the ADH model creates a synthetic control that consists of every donor in the pool. The large RMSPE mentioned above is easier to understand when we see previously unused states, such as Idaho and North Carolina, receiving more than one-quarter of the total weight between them.

To demonstrate the importance of lagged outcome variables as predictors, we also examine the case when only other predictors are used. Figure 7 shows that a model using the four other predictor variables without outcome lags creates a poorly fitted synthetic state in the pretreatment period that is similar to one created by naïvely averaging the 38 donor pool states' outcomes. In table 7, we see that using the four other predictors without any lags includes just three donor states in the synthetic, but Nevada and Utah are omitted. Dropping one of these four predictors from the model without any lags, we find that every donor pool state is included in the synthetic, a sign that the model could not differentiate well between the possible donor states to select those that best represented California had Proposition 99 not been enacted. As expected, the RMSPE in these cases is quite high, ranging from 9.1 to 16.3.

FIGURE 7

Synthetic California Per Capita Cigarette Sales without Outcome Lags

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).

When we have useful predictors, we can find key outcome lags responsible for a synthetic California that closely fits the actual California. If all outcome lags are used, the other predictors are not used. The selection of states varies somewhat as the outcome lags vary but are generally similar. Starting from a set of three lags and varying the other predictors sharply changes the fit of the synthetic California and the selection of donor states. Together, these sections show that when using nonlag predictors, those variables must be carefully chosen to reduce the RMSPE and to create a small set of donor states that remain even if fewer outcome lags are used.

The Year Range for Averaging Predictors

We also examine the results' sensitivity to variation in the range of years in the pretreatment period over which the nonlag predictor variables are averaged. The predictors should be averaged over the pretreatment year period formed from the available data. If the relationship between the outcome variable and the predictors is not stable in sample's early years or if the relationship in the treated state differs from those in the donor states, the early years should not be used to create the averages. Implementing a test of this sort can help SCM practitioners confirm that their predictor year range is long enough to generate a close pretreatment fit and predict posttreatment outcomes. ADH's dataset goes back to 1970,[5] and though their earliest outcome lag is from 1975, their other predictor variables are averaged over 1980–88.[6] We vary this period over which the nonlag predictors are averaged, expanding the range to 1970–88 and 1975–88 and condensing the period to 1984–88. Varying the predictor year range had little effect visually on the synthetic's fit in the pretreatment years (figure 8), but the RMSPE from the condensed 1984–88 predictor period is about 20 to 25 percent higher than that of the other periods (table 4). The higher RMSPE under the 1984–88 model indicates that averaging predictors over the five-year window may not be sufficient and that extending the period will yield a better-fitting synthetic. We find similar problems with the 1984–88 model when we look at the treatment outcome path and donor state selection. In the years following the treatment, the outcome path of the 1984–88 synthetic is slightly lower than that of the other synthetics beginning in 1992. Table 8 shows that the same five states are selected to form the synthetic California for each selected predictor year range except in the 1984–88 model, where Montana drops out. The weighting for Connecticut, Nevada, and Utah remains similar across all the models, but shifts away from Montana and more heavily to Colorado as the predictor period shortens.

TABLE 8

Synthetic California State Weights for Various Predictor Year Ranges

	1970–88	1975–88	1980–88 (ADH)	1984–88
Alabama	0	0	0	0
Arkansas	0	0	0	0
Colorado	0.069	0.063	0.161	0.274
Connecticut	0.105	0.109	0.068	0.087
Delaware	0	0	0	0
Georgia	0	0	0	0
Idaho	0	0	0	0
Illinois	0	0	0	0
Indiana	0	0	0	0
Iowa	0	0	0	0
Kansas	0	0	0	0
Kentucky	0	0	0	0
Louisiana	0	0	0	0
Maine	0	0	0	0
Minnesota	0	0	0	0
Mississippi	0	0	0	0
Missouri	0	0	0	0
Montana	0.231	0.232	0.201	0
Nebraska	0	0	0	0
Nevada	0.254	0.255	0.235	0.255
New Hampshire	0	0	0	0
New Mexico	0	0	0	0
North Carolina	0	0	0	0
North Dakota	0	0	0	0
Ohio	0	0	0	0
Oklahoma	0	0	0	0
Pennsylvania	0	0	0	0
Rhode Island	0	0	0	0
South Carolina	0	0	0	0
South Dakota	0	0	0	0
Tennessee	0	0	0	0
Texas	0	0	0	0
Utah	0.342	0.342	0.335	0.384
Vermont	0	0	0	0
Virginia	0	0	0	0
West Virginia	0	0	0	0
Wisconsin	0	0	0	0
Wyoming	0	0	0	0
Sum	**1.000**	**1.000**	**1.000**	**1.000**

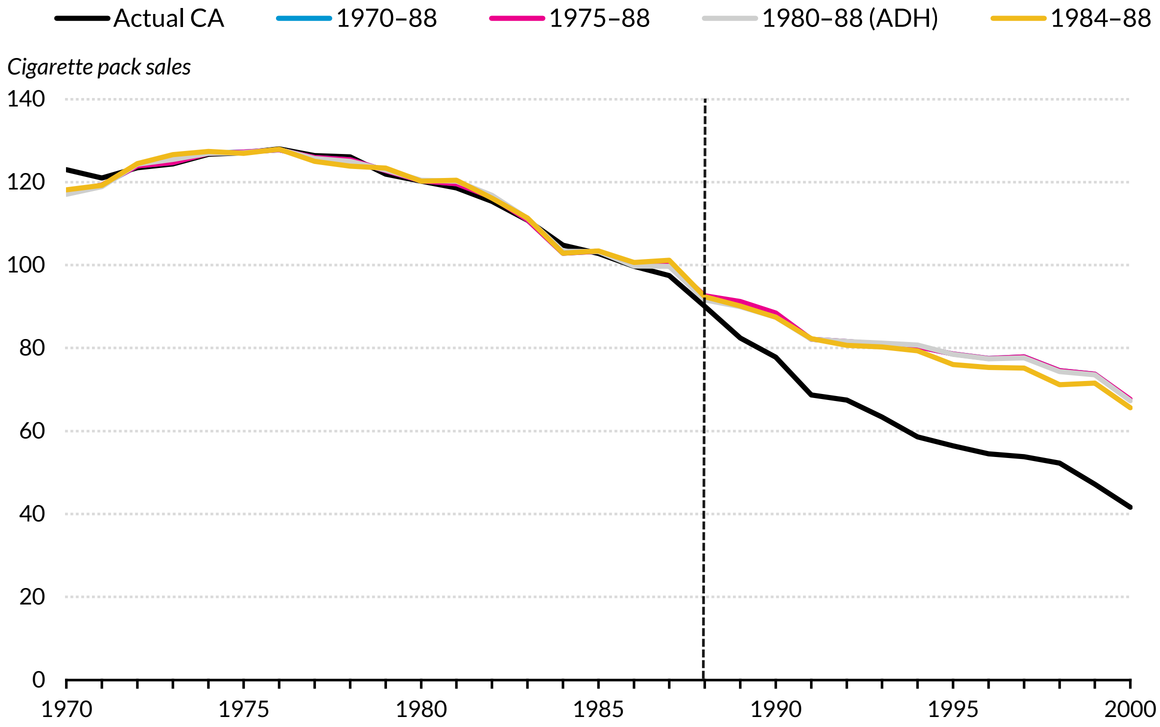
Source: Authors' calculations using the *Synth* package and data from ADH.

Note: ADH = Abadie, Diamond, and Hainmueller (2010).

FIGURE 8

Synthetic California Per Capita Cigarette Sales with Various Predictor Year Ranges

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from ADH.

Notes: ADH = Abadie, Diamond, and Hainmueller (2010). The outcome path of the 1970–88 predictor year range model is almost identical to that of the 1975–88 model, which obscures it in this figure.

For the ADH data, choosing the period over which to average the predictor values is less critical than choosing the outcome lags and other predictors. But in principle, including too many years can lead to a poor fit by the synthetic state, and in ADH's model, using a short predictor period reduces the synthetic California's goodness of fit.

Selecting Predictor Weights

We examine the cross-validation method for selecting predictor weights described in Abadie, Diamond, and Hainmueller (2015). Klößner and coauthors (2016) find that if the number of predictors exceeds the selected donor states when using this method, the outcome path can depend on seemingly meaningless differences, such as the order states are listed. Nevertheless, we evaluate the cross-validation method using our three sensitivity criteria for the resulting synthetic control: the pretreatment fit, the treatment outcome, and the selection of donor states. The method divides the pretreatment period into two ranges: a training period and a validation period. Based on the predictor values in the training period, the predictor weights are selected to minimize the MSPE in the validation period. Those weights are then used with the predictor data in the validation period to create the synthetic control. We compare ADH's standard method with this cross-validation method, dividing the pretreatment period into a 1970–79 training period and a 1980–88 validation period.[7] The cross-validation model we present includes three outcome lags (for 1970, 1975, and 1979) and averages the other predictor variables over the 1970–79 training period. The standard model includes three outcome lags (for 1975, 1980, and 1988) and averages the other predictors over the full 1970–88 pretreatment period (rather than the 1980–88 year range used in ADH's original model). Figure 9 shows that in the pretreatment period, the standard and cross-validation methods yield similar outcome paths, with the standard method leading to a slightly tighter fit and a lower RMSPE (table 4). The cross-validation synthetic begins to stray below the path of the standard synthetic in the final years of the treatment period. The cross-validation synthetic estimates a slightly lower fall in cigarette sales per capita by 2000 because of Proposition 99 (23 packs) compared to the standard model estimate (26 packs). The state weights from the resulting synthetics differ somewhat across the two models (table 9). Colorado and Montana drop out of the cross-validation model, and Utah receives a smaller weight while New Mexico—a new donor state—receives the highest weight (34 percent). While the cross-validation method has appeal as a robustness check for synthetic control results under the standard method, practitioners should be cautious. The concerns raised by Klößner and coauthors (2016) warrant further study and potential adjustments to this method of selecting predictor weights.

TABLE 9

Synthetic California State Weights for Standard Method versus Cross-Validation

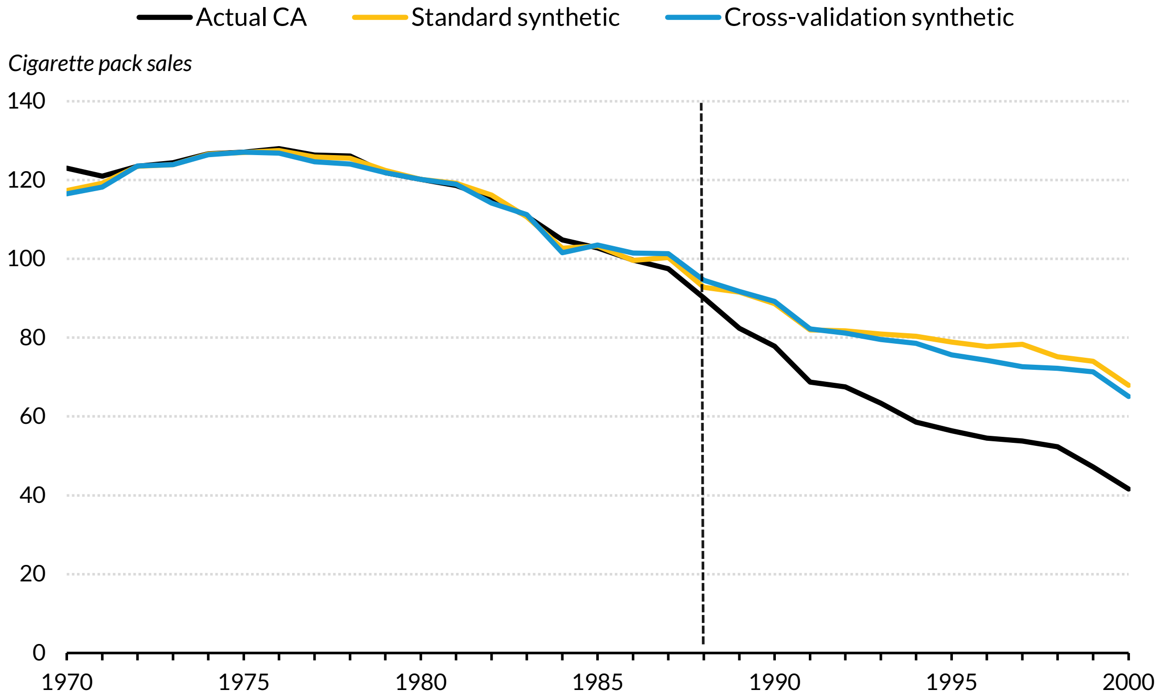
	Standard	Cross-validation
Alabama	0	0
Arkansas	0	0
Colorado	0.007	0
Connecticut	0.126	0.142
Delaware	0	0
Georgia	0	0
Idaho	0	0.072
Illinois	0	0
Indiana	0	0
Iowa	0	0
Kansas	0	0
Kentucky	0	0
Louisiana	0	0
Maine	0	0
Minnesota	0	0
Mississippi	0	0
Missouri	0	0
Montana	0.277	0
Nebraska	0	0
Nevada	0.257	0.260
New Hampshire	0	0
New Mexico	0	0.343
North Carolina	0	0
North Dakota	0	0
Ohio	0	0
Oklahoma	0	0
Pennsylvania	0	0
Rhode Island	0	0
South Carolina	0	0
South Dakota	0	0
Tennessee	0	0
Texas	0	0
Utah	0.333	0.183
Vermont	0	0
Virginia	0	0
West Virginia	0	0
Wisconsin	0	0
Wyoming	0	0
Sum	**1.000**	**1.000**

Source: Authors' calculations using the *Synth* package and data from Abadie, Diamond, and Haimueller (2010).

FIGURE 9

Synthetic California Per Capita Cigarette Sales with Standard Method versus Cross-Validation

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).

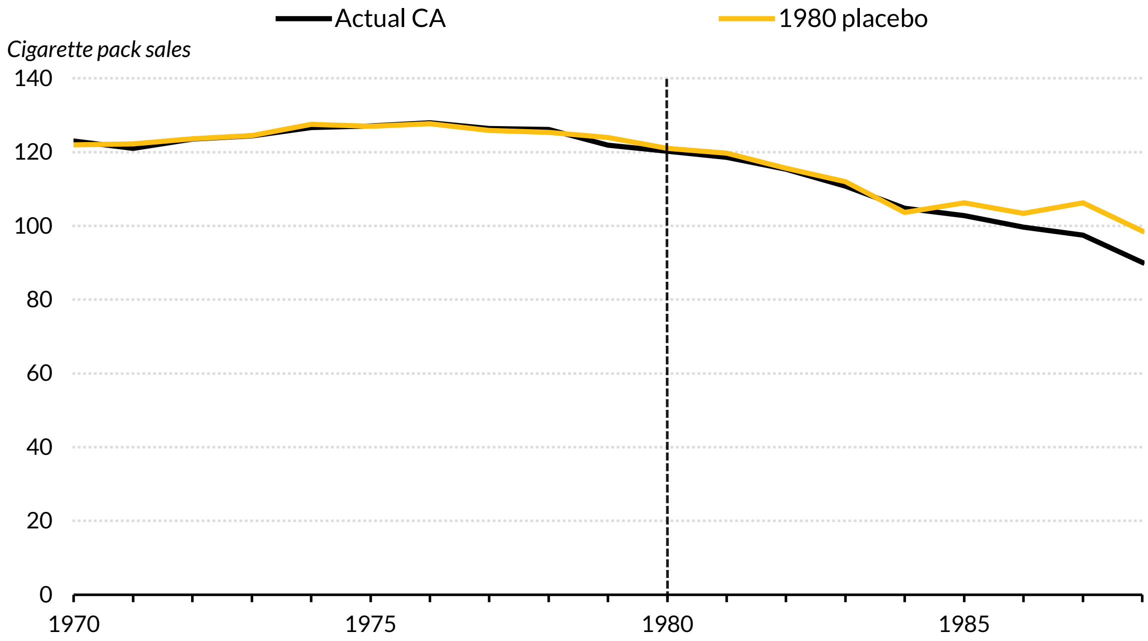
Abadie, Diamond, and Hainmueller's In-Time Placebo and Leave-One-Out Tests

Finally, we examine two additional sensitivity tests described in Abadie, Diamond, and Hainmueller (2015). The first is the "in-time placebo" test, in which they reassign the treatment to occur during the pretreatment period. A placebo estimate differing significantly from the actual pretreatment path would call the model's predictive power into question. We conduct an in-time placebo test using the ADH dataset, where we assign the treatment to the year 1980, roughly in the middle of our 1970–88 pretreatment period.[8] The sample period for this placebo model must end by the year that the actual treatment occurred (1988) to avoid capturing its effects. We use the same predictor variables and years as the cross-validation model in the previous section, including outcome lags for 1970, 1975, and 1979. However, we minimize the MSPE over the 1970–79 period before the 1980 placebo treatment rather than the 1980–88 validation period used in the cross-validation model. Our synthetic California for a 1980 placebo treatment closely follows the path of actual California during the pretreatment period, though there is a divergence in the final four years leading up to the actual treatment (figure 10). At least part of the divergence is because of Connecticut's relatively higher weight in the 1980 placebo model and the zero weight assigned to Montana (table 10). From figure 3, we know that the path of Connecticut cigarette sales per capita did not drift down with California, Nevada, and Utah. Connecticut's higher weight relative to ADH's original model could be the result of the necessary shift in the year range over which predictors are averaged in the 1980 placebo model. We evaluate that possibility by estimating the 1980 placebo test using all available outcome lags before the nonexistent 1980 treatment and no other predictors but find similar state weights (table 10) and outcome paths for both 1980 placebo treatment models. It is likely, therefore, that Connecticut's higher weight is because the 1980 placebo model uses outcome lags for 1970, 1975, and 1979 (rather than 1975, 1980, and 1988 as in ADH's original model) and minimizes the MSPE over the early 1970–79 period.

FIGURE 10

Synthetic California Per Capita Cigarette Sales for a 1980 Placebo Treatment

Before and after a nonexistent treatment in 1980



Source: Authors' estimations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).

TABLE 10

Synthetic California State Weights for a 1980 Placebo Treatment

	1980 test	1980 test with all lags 1970–79
Alabama	0	0
Arkansas	0	0
Colorado	0	0
Connecticut	0.346	0.330
Delaware	0	0
Georgia	0	0
Idaho	0	0
Illinois	0	0
Indiana	0	0
Iowa	0	0
Kansas	0	0
Kentucky	0	0
Louisiana	0	0
Maine	0	0
Minnesota	0	0
Mississippi	0	0
Missouri	0	0
Montana	0	0
Nebraska	0	0
Nevada	0.303	0.283
New Hampshire	0	0
New Mexico	0	0
North Carolina	0	0
North Dakota	0	0
Ohio	0	0
Oklahoma	0	0
Pennsylvania	0	0
Rhode Island	0	0
South Carolina	0	0
South Dakota	0	0
Tennessee	0	0
Texas	0	0
Utah	0.352	0.323
Vermont	0	0
Virginia	0	0
West Virginia	0	0.064
Wisconsin	0	0
Wyoming	0	0
Sum	**1.000**	**1.000**

Source: Authors' calculations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).

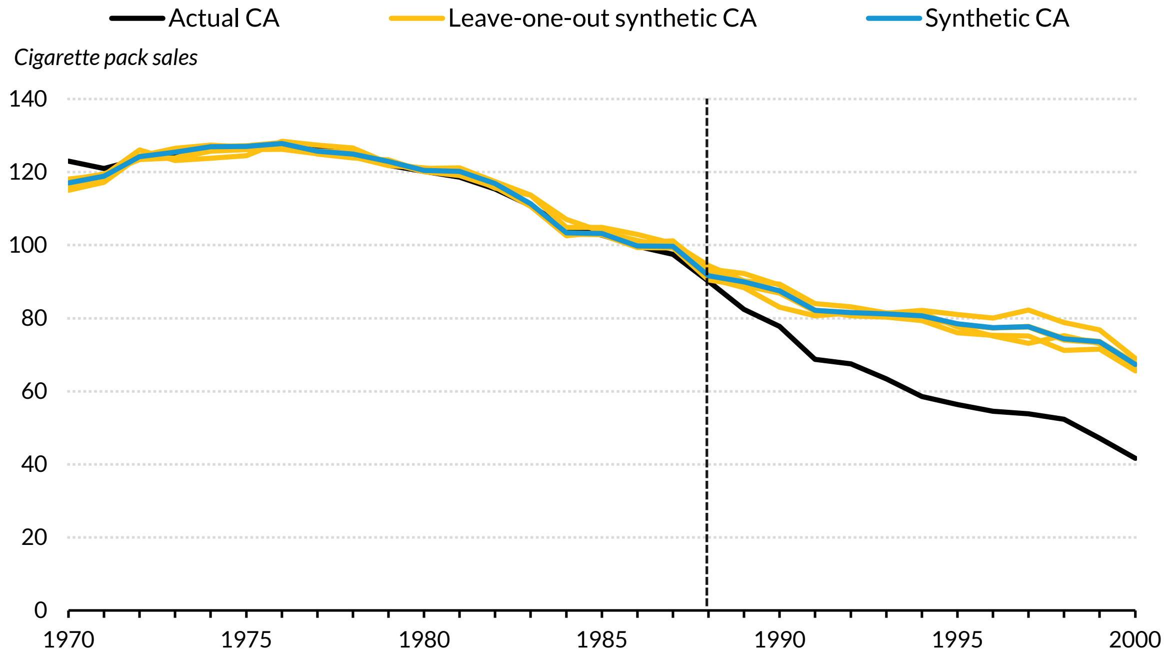
When Proposition 99 was passed in 1988, California's had per capita cigarette sales per capita of 90 packs. The synthetic from ADH's model based on the actual 1988 treatment closely matches this, estimating sales of 92 packs per capita in 1988, while the 1980 placebo model's estimate is slightly higher at 99 packs per capita. But the difference in cigarette sales per capita between the actual and synthetic Californias for a 1980 placebo treatment does not rise above nine packs in the years before Proposition 99 went into effect. Compared to the divergence in the synthetic and actual Californias under ADH's actual 1988 treatment model—which exceeds nine packs in every posttreatment year (1989–2000) except the first—the 1980 placebo's divergence in the final years leading up to 1988 is relatively small. Thus, the in-time placebo test does not demonstrate that the difference between the synthetic and actual Californias in ADH's model arises for reasons other than the treatment.

A second sensitivity test introduced in Abadie, Diamond, and Hainmueller (2015) is the leave-one-out test. Here, they iterate over the model to leave out one selected donor region each time to assess whether one of the donor regions is driving the results. For example, Nevada is used in the synthetic California in the ADH results and in many sensitivity tests. If California citizens responded to the tobacco control program by purchasing cigarettes in nearby Nevada, Proposition 99's benefits would be exaggerated because cigarette purchases shifted to Nevada, forcing the synthetic California to diverge from the actual California. In addition, if one state is pivotal to the results, it may be useful to verify that the state's outcome variable path is credible. For example, the results would be less credible if cigarettes sales per capita rise because of a dramatic drop in the price of cigarettes in that state. We conduct a leave-one-out test with the ADH data, leaving out one of the five donor states (Colorado, Connecticut, Montana, Nevada, and Utah) from the donor pool and repeating the process, generating five additional synthetic controls to compare with the original. Figure 11 shows that the leave-one-out synthetics closely match the original synthetic California that includes all five donor states, verifying the robustness of the original synthetic from ADH. Thus, the results do not appear to be driven by spillover sales in Nevada or any other donor state. This supports one of our necessary assumptions: that California was the only state affected by Proposition 99.

FIGURE 11

Leave-One-Out Test Distribution of Synthetic California Per Capita Cigarette Sales

Before and after Proposition 99 passage in 1988



Source: Authors' estimations using the *Synth* package and data from Abadie, Diamond, and Hainmueller (2010).

Conclusion

The SCM is a quantitative method for assessing regional development policies that can complement more qualitative methods, such as the case study approach. The method has several advantages. It is ideal for examining a policy unique to a particular region. It makes less restrictive assumptions than some other quantitative methods, such as difference-in-differences. The code is freely distributed and straightforward to use. Standard output includes a list of regions and their contribution to the synthetic control region, allowing analysts to make informed judgments about donor regions' comparability with the treated region.

But the method has limitations and should be used with those limitations in mind. Only the region being studied should receive the treatment. There cannot be spillovers, and regions in the donor pool cannot have engaged in similar treatments. The policy cannot have affected the region before going into effect (although violations can be addressed by changing the date when the treatment is presumed to take place). The treated region cannot be an outlier in the pretreatment period. If no donor pool regions have higher and lower levels of the variables used to match the treated region before the policy change, the resulting synthetic control region can be misleading. In the pretreatment period, regions in the donor pool must have comparable predictor variable characteristics with those of the policy region, and those variables must have an approximately linear effect on the outcome. Selecting regions from the donor pool can be sensitive to the choice of variables used to match the donor regions to the treated region. In addition, a seemingly reasonable model choice—matching donor regions to the treated region with outcome lags from all possible years of the pretreatment period—renders other covariates useless. With these cautions in mind, the SCM can be a useful addition to an analyst’s toolkit.

Appendix A. Economic Development Applications

The SCM has been used to analyze many issues. Here, we briefly review two studies of regional development policies. Both studies use the SCM carefully, conducting sensitivity tests or exploring the data in greater depth when necessary.

Castillo and Coauthors' Analysis of Salta's Tourism Development Policy

Castillo and coauthors (2015) examined the effect of the Tourism Development Policy (TDP) in the province of Salta, Argentina, from 2003 through 2010. The authors found that the TDP raised annual employment in the tourism industry about 11 percent. They also conducted sensitivity tests and concluded that their results are robust to the potential objections they raised.

The authors chose 11 pretreatment covariates: employment level, number of firms, average monthly wage, average number of employees per firm, average age of firms, annual GDP, share of employees ages 18 and older without pension contributions, log of population ages 14 and older, share of population ages 20 and older with completed university education, share of households with access to paved roads in the census area, and share of households with access to public lighting in the census area. The authors did not mention which lags were used.

Several provinces were eliminated from the donor pool because they underwent structural shocks to tourism employment or because they had substantially different values of the covariates or a different structural process driving their outcomes. This left 19 provinces in the donor pool. The SCM chose five provinces: Formosa, Jujuy, Neuquén, Santa Fe, and Tucumán. Their outcome variable was employment in the hospitality sector, and they used monthly sector-level panel data from 1996 to 2013.

The authors also created two alternative synthetic Saltas. The first was made up of other sectors of Salta's economy rather than the tourism sector from other provinces. Identification in that alternative depends on the lack of economy-wide policy changes in Salta that would affect the control sectors of the economy and the tourism sector. The second alternative used other economic sectors in other provinces to create a synthetic control for Salta's tourism sector, providing 900 sector/province units.

The results from comparing employment in the hospitality sector of the actual Salta to employment in each of the synthetic Saltas collectively suggested that the TDP raised tourism employment, but these results were not entirely convincing. First, the synthetic Salta built from other provinces may have been biased because of spillover effects. Second, the results may have been because of chance. Third, the results may have been because of a general improvement in Salta's economy around the time of the TDP.

The first objection arose because the state of Jujuy received nearly 40 percent of the weight in the synthetic Salta formed from other provinces. In the second alternative, the synthetic Salta also consisted largely of sectors from Jujuy. That reliance, plus the long shared border between Salta and Jujuy, raised the possibility that spillover effects biased the results. Positive spillover would indicate the synthetic Salta partially incorporated the employment boost provided by the TDP, and the results underestimated its effect. Negative spillover would indicate the TDP's effect was instead overestimated. The authors used a leave-one-out test to demonstrate that omitting any state, including Jujuy, did not substantially change the results.

The remaining objections were addressed through placebo tests. State-level placebo test results showed that the gap between the actual and synthetic Saltas was larger than the difference created using any other province. That suggests that the results were not solely because of chance. The results of a Salta province sector-level placebo test showed that the actual-to-synthetic gap in the tourism sector was much larger than for any other sector, suggesting that the results were not because of changes in Salta's overall economy. The results of an in-time placebo test suggest that nothing before 2003 caused the divergence between the actual and synthetic Saltas.

Ando's Analysis of Japanese Nuclear Power Facilities

Ando (2015) separately examined the change in real per capita taxable income associated with siting eight nuclear power facilities (NPF) in Japanese municipalities from 1975 through 1988. Many believe that an NPF provides numerous jobs and fiscal benefits to municipalities, such as grants from the central government and increased revenues from property taxes. Ando concluded that NPF siting had strong effects in two instances, smaller effects in three others, and no significant effect in the final three cases. He investigated the source of the variation and concluded that the municipalities most strongly affected saw increased employment in the construction industry.

For predictors, the author chose real per capital taxable income, the employment rate, an urbanization index, the share of the population ages 16 to 64, the share of the population ages 65 and over, the share of employment in various economic sectors, and population growth rates for several age groups. Seven of the sample periods began in 1972, and one began in 1981; all ended in 2002.

For the eight analyses, the donor pool was limited to municipalities in the same region as, but not sharing a border with, the municipality receiving the NPF. Because the NPFs were all located in coastal areas, noncoastal municipalities were also eliminated from the donor pool. Possibly because each of the eight analyses had separate donor pools (385 municipalities in total), the municipalities used to form the synthetic controls were not listed.

The results varied considerably across the eight municipalities. Two cities, Rokkasho and Tomioka, experienced about a 25 percent increase in average taxable income. Other cities, such as Onagawa, saw much smaller increases. Placebo tests supported the hypothesis that Rokkasho and Tomioka saw increases in average taxable income, but provided weaker evidence of increased incomes in Tomari, Naraha, and Kariwa. The outcome paths of Onagawa, Kashiwazaki, and Shika were indistinguishable from the paths of placebo cities. Comparing the average outcomes for the eight NPF municipalities with a histogram of the average placebo outcomes yielded similar evidence, although Ando found no compelling evidence that Naraha saw an increase in average income.

To explore such strong variation in outcomes, Ando examined employment trends in seven economic sectors. He concluded that construction and manufacturing employment in Rokkasho and Tomioka responded more strongly than they did in the other cities. He also examined the effect on per capita public spending and average land prices, concluding that increased public spending did not account for the variation in outcomes but that increased land costs in several municipalities partially explained the small increase in per capita income. The need to answer why some cities saw more responsive construction and manufacturing employment and others saw sharp increases in land prices demonstrates that the SCM supplements rather than replaces the case study approach.

Appendix B. *Synth* Package Predictor Weight Selection

By default, the *Synth* code uses the regression-based method to select predictor weights. This is the fastest method in the *Synth* package, selecting a synthetic composed of the best-fitting state weights conditional on the regression-based V-weight matrix. Specifying the "nested" option in the code will result in using the fully nested optimization method, a lengthier process yielding more precise estimates.[9] The nested specification begins with the regression-based method, but produces other combinations of predictor weights to achieve the lowest root mean squared prediction error. We use the nested specification in all the estimates we report to maintain consistency with ADH's work. Comparing the two methods, we find that while the units of the predictor variables are of no consequence when using the default regression-based method, converting to units of different magnitudes (e.g., describing the price of a package of cigarettes in dollars rather than cents) while using the nested specification has small effects on the state weights. This may merit future adjustments to the nested option. Alternatively, users can specify their own predictor weights using the "customV" option.

Notes

1. Jens Hainmueller, “Synth Package,” Stanford University, accessed January 13, 2017, https://web.stanford.edu/~jhain/synthpage.html.
2. “SYNTH: Stata module to implement Synthetic Control Methods for Comparative Case Studies,” Örebro University Business School, last updated January 8, 2017, http://econpapers.repec.org/software/bocbocode/S457334.htm.
3. In principle, a weighted sum of outcomes from donor states could equal the treated state if the weights were not bound by the unit interval. But using those weights could create a synthetic control that is not representative of the treated state in the pretreatment period.
4. Accessed March 24, 2016.
5. The log of per capita GDP is only available in the data beginning in 1972, and per capita beer consumption is only available beginning in 1984.
6. This does not include per capita beer consumption, which is averaged over 1984–88.
7. To keep a clean comparison, we do not include per capita beer consumption in our predictor variables for either model. We do not have this variable in the data until 1984, so we could not average it over the 1970–79 training period in the cross-validation model.
8. Again, we do not include per capita beer consumption in our predictor variables, as it is not available until 1984.
9. A second option called “allopt” can be specified with the nested option to produce the most robust estimates. This specification will increase computing time.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105 (490): 493–505.

———. 2015. “Comparative Politics and the Synthetic Control Method.” *American Journal of Political Science* 59 (2): 495–510.

Abadie, Alberto, and Javier Gardeazabal. 2003. “The Economic Costs of Conflict: A Case Study of the Basque Country.” *American Economic Review* 93 (1): 112–32.

Ando, Michihito. 2015. “Dreams of Urbanization: Quantitative Case Studies on the Local Impacts of Nuclear Power Facilities Using the Synthetic Control Method.” *Journal of Urban Economics* 85: 68–85.

Athey, Susan, and Guido W. Imbens. 2006. “Identification and Inference in Nonlinear Difference-in-Difference Models.” *Econometrica* 74 (2): 431–97.

Castillo, Victoria, Lucas Figal Garone, Alessandro Maffioli, and Lina Salazar. 2015. “Tourism Policy, a Big Push to Employment? Evidence from a Multiple Synthetic Control Approach.” Inter-American Development Bank Working Paper Series, IDB-WP-572. Washington, DC: Inter-American Development Bank.

Ferman, Bruno and Cristine Pinto. 2016. “Revisiting the Synthetic Control Estimator.” MPRA Working Paper No. 75128. Munich, Germany: Munich University Library, Munich Personal RePEc Archive.

Ferman, Bruno, Cristine Pinto, and Vitor Possebom. 2016. “Cherry Picking with Synthetic Controls.” FGV Working Paper 420. São Paulo, Brazil: Sao Paulo School of Economics.

Galiani, Sebastian, and Brian Quistorff. 2016. “The *synth_runner* Package: Utilities to Automate Synthetic Control Estimation Using *synth*.” College Park: University of Maryland.

Glantz, Stanton A., and Edith D. Balbach. 2000. *Tobacco War: Inside the California Battles*. Berkeley: University of California Press.

Hahn, Jinyong and Ruoyao Shi. 2016. “Synthetic Control and Inference.” Working paper.

Kaul, Ashok, Stefan Klößner, Gregor Pfeifer, and Manuel Schieler. 2016. “Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors.” Working Paper. Saarbrücken, Germany: Saarland University.

Klößner, Stefan, Ashok Kaul, Gregor Pfeifer, and Manuel Schieler. 2017. “Comparative Politics and the Synthetic Control Method Revisited: A Note on Abadie et al. (2015).” Working Paper. Saarbrücken, Germany: Saarland University.

About the Authors

Robert McClelland is a senior fellow in the Urban-Brookings Tax Policy Center. His work focuses on the behavioral effects of taxation. He received a BA in economics and environmental studies from the University of Santa Cruz and a PhD in economics from the University of California, Davis.

Sarah Gault is a research associate in the Urban-Brookings Tax Policy Center, contributing to the State and Local Finance Initiative. She works primarily on topics relating to state and local public finance, and she has supported research on financial transaction taxes and simplifying student financial aid. Gault holds a bachelor's degree in economics from the College of William and Mary.

Statement of Independence